Which Intermediary Costs Matter for Asset Prices?*

Manav Chaudhary† Julie Zhiyu Fu‡ Jian Li§

March 26, 2026

Abstract

When intermediaries are constrained, asset prices can become distorted. Arbitrage spreads are widely used to diagnose such distortions, but do they reveal distortions in the level of asset prices—the prices that govern borrowing costs, discount rates, and the transmission of monetary policy? We develop a model where intermediaries face two costs: costs proportional to portfolio risk and gross position costs tied to position size. We show that gross position costs predominantly drive spread distortions, while risk-based costs primarily drive price level distortions. The theory delivers a sufficient statistic approach: differences in the rate at which yields and spreads revert after a demand shock separately identify these costs. We take this to the U.S. Treasury cash and OIS markets, exploiting high-frequency shocks from Treasury auctions. On average, risk-based costs dominate: both Treasury yields and OIS rates respond strongly to demand shocks but comove closely, leaving spreads largely unchanged, so a spread-based diagnostic misses the dominant source of price-level distortions. A calibrated model reinforces the disconnect: relaxing position-size constraints such as the supplementary leverage ratio substantially reduces spread volatility with little effect on yield volatility, while reducing risk-based regulatory requirements has the opposite relative effect.

*We thank Daniel Barth, Ian Martin, Kathy Yuan, Peter Kondor, Dimitri Vayanos and seminar participants at the Demand in Asset Markets workshop, London Junior Conference, London School of Economics and WAPFIN for helpful comments.

†London School of Economics. Email: m.chaudhary6@lse.ac.uk

‡Washington University in St. Louis, Olin Business School. Email: z.fu@wustl.edu

§Columbia Business School. Email: jl5964@columbia.edu

1 Introduction

Intermediaries, such as broker-dealers and hedge funds, play a central role in determining asset prices across capital markets. When these intermediaries are constrained, prices can become distorted, impairing capital allocation, risk sharing, and liquidity provision.

Arbitrage spreads—the price difference between two assets that deliver identical payoffs—are a standard way to assess if intermediary constraints are distorting prices. Absent frictions, intermediaries should drive these spreads to zero. When a riskless arbitrage relationship "breaks," it is taken as a sign that intermediation capacity has become scarce and that liquidity provision is impaired. Central banks treat these spreads as barometers of market functioning, and a large empirical literature traces the effects of intermediation capacity through the behavior of these spreads (Du, Tepper, and Verdelhan, 2018; Du, Hébert, and Huber, 2023; D. Duffie et al., 2023).[1]

But do arbitrage spreads provide an accurate picture of distortions in the *level* of asset prices—the distortions that govern borrowing costs, discount rates, and the transmission of monetary policy? Theory does not provide a clear-cut answer. In canonical intermediary asset pricing models, intermediaries face risk-based constraints that are first-order drivers of price levels (He and Krishnamurthy, 2013; Brunnermeier and Sannikov, 2014; Vayanos and Vila, 2021), but these constraints do not generate arbitrage violations, so spreads remain silent even when price levels are substantially distorted. Models of arbitrage violations show that spreads arise from gross position constraints such as leverage ratios and margin requirements (Gromb and Vayanos, 2002; Gârleanu and Pedersen, 2011; Du, Tepper, and Verdelhan, 2018), but do not address what these constraints imply for price levels. Answering this question requires a framework that jointly models and quantifies how different types of intermediary constraints affect both spreads and price levels.

This paper develops such a framework. We build a Vayanos and Vila, 2021 style model with slow-moving capital in which intermediaries arbitrage across a cash and synthetic asset with identical payoffs. Intermediaries face two broad costs. The first is a risk-based cost penalizing portfolio risk, as would arise from capital adequacy requirements or value-at-risk limits. The second is a gross position cost tied to the size of positions, as would arise from the supplementary leverage ratio or margin requirements on individual positions. The model delivers a sharp distinction between the two. While gross position costs matter primarily for

[1] The Federal Reserve's expansion of Treasury purchases in March 2020 was prompted in part by the breakdown of the Treasury cash-futures basis, and the Bank of England's temporary gilt purchase programme in September 2022 was triggered by sharp dislocations in gilt spreads. In both cases, subsequent spread compression was cited as evidence that the interventions had restored market functioning. See Board of Governors of the Federal Reserve System (2020) and Bank of England (2025) for more details.

arbitrage spreads, risk-based costs are the dominant force behind movements in price levels. The intuition is that trading against price-level deviations requires directional positions that expose intermediaries to portfolio risk, while arbitraging the spread involves hedged positions that are costly in gross balance sheet usage but largely neutral in risk. Absent gross position costs, the model admits an equilibrium where two identical assets are perfectly integrated and the spread is always zero, highlighting the tight link between spreads and position costs. This distinction underscores the importance of separately identifying the two types of costs.

We develop a sufficient statistic approach to quantify the relative importance of these two costs. An exogenous demand shock generates an immediate price impact that then decays as intermediaries unwind their positions as they sell to slow-moving capital. The intermediaries' first-order condition links the initial rate of this decay to the marginal cost of intermediation—the initial decay rate is a sufficient statistic for intermediary costs. Because the two costs affect price levels and spreads differently, they induce distinct decay rates in yields and spreads, allowing us to separately identify each type of cost.

We take this approach to the U.S. Treasury cash market and the overnight index swap (OIS) market, using high-frequency identified demand shocks around Treasury auction result releases (Droste, Gorodnichenko, and Ray, 2024). In the full sample, we find that risk-based costs are, on average, the dominant driver of intermediation capacity. Following a demand shock, Treasury yields and the OIS rate move closely together, leaving the Treasury–OIS spread largely unchanged. As a result, a diagnostic based solely on spreads would miss the risk-based costs that primarily govern price-level dynamics. These findings hold unconditionally, but the relative importance of the two costs shifts during crises. During the 2008 Global Financial Crisis, risk-based costs are the dominant driver of intermediation capacity; during the COVID crisis, gross position costs are. This shift is consistent with the post-2017 implementation of the Supplementary Leverage Ratio (SLR), which introduced significant gross position costs for broker-dealers in the Treasury market. However, these costs bind only in periods of acute market stress. Outside of crisis episodes, risk-based costs remain the dominant driver of intermediation capacity even in the post-2017 sample, and spreads remain a limited diagnostic.

We describe the model in more detail. We model two assets with identical fundamental values, traded by three types of agents: institutional investors such as pension funds and asset managers, intermediaries such as dealers and hedge funds, and noise traders. Institutional investors are long-term investors with downward-sloping demand that depends on the assets' prices and their fundamental value. To capture the slow adjustments in their positions, we assume they can only partially adjust their position towards the target demand each period.

Intermediaries act as arbitrageurs between the two markets, facing two types of costs: a risk-based cost for bearing portfolio risk, and a gross position cost that captures constraints tied to the size of asset positions. Finally, noise traders trade for exogenous reasons and generate demand shocks.

We solve the equilibrium in closed form, where the prices are linear functions of the intermediaries' inventories, noise trader demand, and the fundamental value of the asset. To highlight the importance of gross position costs, we show that absent such costs, the two asset markets can comove perfectly, in which case the spread is always zero. Our result higlights the tight link between spreads and position costs. Furthermore, we characterize the dynamics of prices and spreads following an exogenous demand shock. Intermediaries act as the marginal, short-horizon liquidity providers: they absorb the shock on impact and then gradually unwind their inventories as slow-moving institutional investors absorb the position over time.

As a result, the impulse response of prices naturally decomposes into two components. The long-run response is pinned down by institutional investors' demand elasticities, since they determine the ultimate allocation once the intermediary inventory has been offloaded. The additional short-run transitory response reflects the compensation required by intermediaries to hold inventory while this adjustment is underway. The dynamics of this transitory component motivate our identification strategy.

We use the model to study how different types of intermediation costs shape prices and spreads differentially. Risk-based costs *integrate* the two markets, while gross position costs *segment* them. The former encourages intermediaries to take offsetting positions that hedge risk, linking the two markets through active arbitrage and raising return correlation; the latter penalizes the gross positions required to do so, driving prices apart. This distinction determines when spreads are—and are not—informative about distortions in price levels.

Given their different implications, we attempt to learn about the relative importance of these two types of intermediation costs empirically. The observed price response at any horizon reflects both end-investor demand elasticities and intermediary inventory-bearing costs. Because these two forces are confounded, we cannot identify intermediation costs from the price impact directly. However, we show that the speed at which prices and spreads revert immediately following an exogenous demand shock pins down the intermediation costs. The intuition is as follows. Upon a demand shock, intermediaries are the immediate liquidity providers and absorb the entire demand shock on impact. The initial rate of price adjustment reflects the return on the intermediary's inventory and, by the intermediary's first-order condition, is pinned down by marginal intermediation costs. Similarly, the corresponding

initial adjustment in the spread captures the compensation for largely hedged relative-value positions, compensation that is primarily driven by gross position costs. Hence, combining the slope of the price and spread adjustments immediately following a demand shock allows us to separately identify the two types of intermediation costs. This identification strategy applies broadly: the key assumption is that intermediaries are the marginal liquidity providers on impact, a common feature in many models of intermediation that is consistent with the empirical evidence.

We apply this framework to study the U.S. Treasury cash market and the overnight index swap (OIS) market. Following Droste, Gorodnichenko, and Ray (2024), we identify Treasury demand shocks using the high-frequency price impact of Treasury auction result releases. In U.S. Treasury auctions, the amount to be issued is preannounced, but the investor demand is only revealed when the auction closes and results are published. As a result, yield movements in a tight window around the result release capture unanticipated variation in demand and provide a high-frequency proxy for exogenous demand shocks to the Treasury market.

We estimate impulse responses of yields and spreads to auction shocks using local projections and recover initial decay rates from the estimated impulse responses. The theory framework then allows us to map these initial decay rates to the underlying intermediation costs. Because we do not have a direct measure of the size of the demand shocks, we can only identify the relative importance of the two types of intermediation costs, not their absolute magnitudes.

In the full sample (2008–2022), we find that risk-based costs are, on average, the dominant driver of price-level dynamics. In response to auction shocks, Treasury yields and OIS rates move closely together, while the Treasury–OIS spread exhibits little systematic response, implying that risk-based costs dominate intermediation costs on average. Hence, focusing solely on spreads would tend to understate the extent of illiquidity in the market, as spreads miss much of the relevant price-level variation in the Treasury market.

However, the relative importance of the two costs varies significantly across crises. During the Global Financial Crisis, spread responses remain muted, consistent with risk-based costs being the dominant friction. During the COVID episode, by contrast, the Treasury–OIS spread jumps and then decays, indicating that gross position costs became the dominant friction on intermediation capacity in that period of market stress (J. D. Duffie, 2023), suggesting that in the post-Supplementary Leverage Ratio (SLR) regulatory regime, gross position costs tend to become a dominant driver of intermediation capacity in periods of market stress. As a result, spreads are now more informative about price-level distortions during periods of market stress. However, quantifying the impact of these costs on price

dynamics requires a fully calibrated model.

We calibrate the model by targeting the unconditional variance profiles of Treasury yields and the Treasury-OIS spread and conduct two exercises. First, we apply the calibrated model to the COVID-19 episode to quantify the role of elevated gross position costs. The elevated gross position cost almost doubled the spread's price impact and increased spread volatility by over 250%, compared with much more modest increases for yields. So while spreads directionally reflected increased distortions in the Treasury yield during COVID-19, they overstated the magnitude of yield distortions, as gross position costs during the crisis disproportionately affected spread dynamics relative to yield dynamics.

Second, given the current debate on relaxing banking regulations, we examine how hypothetical reductions in each cost component affect Treasury market functioning. A sharp asymmetry emerges: a reduction in the gross position costs (as would arise from relaxing the SLR) compresses the spread response to demand shocks and spread volatility substantially, while leaving yield dynamics relatively unchanged. A reduction in the risk cost (as would arise from relaxing the Stress Capital Buffer) does the opposite, materially reducing yield price impact with little effect on the spread. These counterfactuals illustrate how different regulatory tools affect Treasury market functioning through distinct channels.

The rest of the paper proceeds as follows. We review the relevant literature in the remainder of this section. Section 2 describes the model and our theoretical results. Section 3 discusses our identification strategy and describes the estimation procedure. Section 4 presents the empirical results. Section 5 calibrates the model and presents counterfactual exercises. Section 6 concludes.

1.1 Literature

We contribute to the intermediary asset pricing literature by addressing two related gaps that have received limited systematic attention: the absence of a unified framework that jointly models how different types of intermediation costs affect prices and spreads, and the disconnect between empirical evidence from arbitrage spreads and implications for asset price levels (Haddad and Muir, 2025). We tackle both with a theoretical framework and an empirical identification strategy that the framework motivates, quantifying the relative importance of different types of intermediation costs.

The theoretical literature models intermediaries as marginal investors and studies how their constraints and frictions affect asset prices. Both risk-based constraints, such as capital adequacy requirements and value-at-risk limits, and gross position costs, such as the supplementary leverage ratio and margin requirements, feature prominently in theoretical

work (e.g., Adrian and Shin, 2014; Gârleanu and Pedersen, 2011; Du, Hébert, and Huber, 2023).[2] Much of the existing work, however, either focuses on a single asset without separately identifying risk-based and gross position costs (e.g., He and Krishnamurthy, 2013; Samuel G. Hanson, Malkhozov, and Venter, 2024), or studies one type of constraint in isolation (e.g., Du, Hébert, and Li, 2023). We provide a two-asset framework that jointly studies the effects of both types of costs on prices and spreads. The model delivers a sharp distinction: risk-based costs disproportionately affect price levels, while gross position costs disproportionately affect spreads.

The empirical literature shows that intermediary leverage comoves with returns and prices risk across asset classes (e.g., Adrian, Etula, and Muir, 2014; He, Kelly, and Manela, 2017; Haddad and Muir, 2021), supporting intermediary asset pricing models. Such evidence, however, is subject to critiques regarding the endogeneity of intermediary leverage (Santos and Veronesi, 2022). The existence of arbitrage spreads in heavily intermediated markets, such as CIP deviations, and their widening when intermediary constraints tighten provide more direct evidence in favor of intermediary frictions (see, for example, Du, Tepper, and Verdelhan, 2018; Barth and Kahn, 2025; Boyarchenko et al., 2018; D. Duffie et al., 2023). The remaining question is quantitative: how informative are spread movements about distortions in price levels? We contribute by separately identifying the two types of costs and showing that risk-based costs are the dominant driver of price-level dynamics, while spreads alone are a limited diagnostic for the extent of intermediation costs.

Our methodology draws on the recent literature focusing on the role of demand shocks in asset pricing. Our model builds on the workhorse framework of Vayanos and Vila (2021), which studies how arbitrageurs absorb demand shocks. Our empirical strategy builds on the key insight from demand-system asset pricing that the price response to a demand shock reveals the underlying frictions faced by the arbitrageur (Koijen and Yogo, 2019; Gabaix and Koijen, 2021). The existing literature typically estimates price impact or demand elasticities in reduced form. We take a step further by interpreting the impulse responses of prices and spreads to a demand shock through the lens of our model, thereby identifying the structural parameters governing risk-based and gross position costs.

[2]Given the breadth of this literature, we do not attempt a comprehensive review here. For recent surveys, see He and Krishnamurthy (2018) and Haddad and Muir (2025).

2 Model

To study the implication of intermediary frictions on asset prices and spreads, we introduce a preferred-habitat style model with two assets (Vayanos and Vila, 2021). We first describe the model setup in detail and analyze the equilibrium. We then discuss how different types of intermediary costs matter for asset prices and their relative spreads.

2.1 Setup

Time is continuous and there are two assets with identical fundamental values. Our model features three types of agents: institutional investors that demand both assets but adjust their capital only sluggishly, an intermediary sector that arbitrages across markets and noise traders in each market.[3]

Assets. There are two assets in the economy: a cash asset denoted by C, such as the Treasury bonds, and a synthetic asset denoted by S, which is a derivative contract that replicates the payoff of the cash asset, such as the OIS swap. The two assets have the same fundamental value V_t, which follows a Brownian motion. Absent any frictions, the prices of the two assets should be identical and equal to V_t, i.e., $P_{C,t} = P_{S,t} = V_t$. We interpret V_t as the long-term fundamental value of the two assets. We are interested in how different types of intermediation costs affect the level of asset prices and the relative spread between the two assets, defined as $P_{S,t} - P_{C,t}$, differently. We use $\mathbf{P}_t$ to denote the 2 by 1 price vector of the two assets.

Institutional Investors. Institutional investors are modeled as having slow moving capital with reduced form demand. These are long-term investors who care about the fundamental payoff of the assets. They have downward sloping demand curves and each period, their positions adjust only partially toward their target allocation (D. Duffie, 2010). These could be mutual funds, pension funds and insurance companies — investors that either face tight investment mandates or lack the sophistication (or incentives) to arbitrage actively across markets. They are also often slow to respond when prices deviate from fundamentals, hence we model them as slow-moving capital, similar to Greenwood, Samuel G Hanson, and Liao (2018).

[3] We separate noise traders from institutional investors for expositional clarity, but their shocks can equally be thought of as arising from institutional investors.

We denote their holdings in the cash and synthetic asset markets by $y_{C,t}$ and $y_{S,t}$, respectively. Each instantaneous period, a fraction kdt of the investors in each market re-optimize their holdings to their target demand $Z_{i,t}$ ($i \in \{C, S\}$), while the rest keep their previous holdings. The target demand curve of the institutional investors who re-optimize in each period is given by

$$\begin{pmatrix} Z_{C,t} \\ Z_{S,t} \end{pmatrix} = -\boldsymbol{\zeta} \begin{pmatrix} P_{C,t} - V_t \\ P_{S,t} - V_t \end{pmatrix} + \theta, \tag{1}$$

where $\boldsymbol{\zeta}$ is the demand elasticity matrix. The target demand of institutional investors is decreasing in the prices and increasing in the fundamental value. The constant term $\boldsymbol{\theta}$ captures an exogenous component of the demand, which captures other unmodelled components of demand such as institutional investors' hedging incentives and flows.

The dynamics of institutional investors' aggregate holdings are hence given by

$$\begin{pmatrix} dy_{C,t} \\ dy_{S,t} \end{pmatrix} = k \begin{pmatrix} Z_{C,t} - y_{C,t} \\ Z_{S,t} - y_{S,t} \end{pmatrix} dt. \tag{2}$$

The institutional investors' holdings as a sector adjust slowly to their targeted demand.

Intermediaries The intermediary sector acts as the arbitraguer between the two markets. They are risk-averse and choose their positions in the two asset markets, $x_{i,t}$ ($i \in \{C, S\}$) to maximize their instantaneous expected utility, subject to two types of intermediation costs: risk-based cost and an additional gross position cost in each asset. More specifically, the intermediary's optimization problem is given by

$$\max_{x_{C,t}, x_{S,t}} \mathbb{E}_t \left[dW_t\right] - \frac{\gamma}{2} Var_t(dW_t) - (\frac{\psi_C}{2} x_{C,t}^2 + \frac{\psi_S}{2} x_{S,t}^2) \tag{3}$$

where

$$dW_t = x_{C,t} dP_{C,t} + x_{S,t} dP_{S,t} + \left(W_t - \sum_{i \in \{C,S\}} x_{i,t} P_{i,t} \right) r dt. \tag{4}$$

The first term captures the expected return of the intermediaries portfolio dW_t. Given they are short-term investors, they do not care about the fundamental value of the assets and only care about the instantaneous return of their portfolio. Here r is the risk-free rate. Given the frequency of our analysis in the empirical section is at the daily level, the risk-free return is

close to zero. Hence we set it to zero in our main theoretical analysis as well. However, all results are robust to the general case.

The second term $\frac{\gamma}{2}Var_t(dW_t)$ captures risk-based cost of the intermediaries, and γ is the risk aversion coefficient. Their risk-based cost is proportional to the riskiness of their entire portfolio, reflecting intermediary costs such as the value-as-risk limits and other types of variance-based constraints. The last term $(\frac{\psi_C}{2}x_{C,t}^2 + \frac{\psi_S}{2}x_{S,t}^2)$ represents the additional cost related to their position in each asset, capturing balance sheet costs that are increasing in the size of the asset positions but do not take into account the rest of the portfolio, such as the Supplementary Leverage Ratio (SLR) constraint. We refer to ψ_C and ψ_S as the gross position cost parameters.

We do not aim to separate all the frictions faced by intermediaries in detail in this paper, but rather to capture the two broad types of intermediation costs that are commonly studied in the literature and cited in policy discussions. The risk-based cost reflects the fact that intermediaries are risk-averse and certain regulations are tied to the overall risk of their portfolio. The position cost captures the fact that intermediaries face constraints in position sizes. Both types of costs can represent multiple types of frictions in practice. Our goal is to understand how these two broad types of costs matter for asset prices and spreads differently, and to provide a way to empirically identify them separately.

Noise Traders. In each market, the noise traders hold $\beta_{i,t}$ ($i \in \{C, S\}$) units of the asset. We use $\boldsymbol{\beta}_t$ to denote the vector form. Then $d\beta_{C,t}$ and $d\beta_{S,t}$ captures the demand shocks that originate from the noise traders. $(\beta_{C,t}, \beta_{S,t}, V_t)^\top$ follows an Ornstein–Uhlenbeck process with mean reversion coefficient $\boldsymbol{\eta} = \text{diag}(\eta_{\beta_C}, \eta_{\beta_S}, \eta_V)$ and potentially correlated shocks.

$$\begin{pmatrix} d\beta_{C,t} \\ d\beta_{S,t} \\ dV_t \end{pmatrix} = \boldsymbol{\eta} \begin{pmatrix} \bar{\beta_C} - \beta_{C,t} \\ \bar{\beta_S} - \beta_{S,t} \\ \bar{V} - V_t \end{pmatrix} dt + \hat{\Sigma}^{\frac{1}{2}} d\mathbf{B}_t \tag{5}$$

where

$$\hat{\Sigma} = \begin{pmatrix} \sigma_C^2 & \sigma_{CS} & \sigma_{CV} \\ \sigma_{CS} & \sigma_S^2 & \sigma_{SV} \\ \sigma_{CV} & \sigma_{SV} & \sigma_V^2 \end{pmatrix} \tag{6}$$

Our general framework allows for flexible mean reversion. However, to have the sharpest characterization of the price dynamics, in the following analysis we consider the asymptotic case where the mean reversion $\boldsymbol{\eta}$ goes to $\mathbf{0}$ so that the noise trader demand and the

fundamental value behave like martingales.

Finally, we assume the asset supply is fixed at $\bar{S}_i$ for $i \in \{C, S\}$. The market clearing condition implies

$$x_{i,t} + y_{i,t} + \beta_{i,t} = \bar{S}_i, \quad i \in \{C, S\}. \tag{7}$$

Due to the quadratic costs in the intermediaries' optimization problem, the equilibrium price is linear in the state variables. We characterize the equilibruim in closed form in the next section.

2.2 Equilibrium Analysis

We characterize the equilibrium price and quantity dynamics in closed form in this section. The proof is in Appendix A. Similar to the Vayanos-Vila framework, we conjecture the equilibrium price of the two assets can be expressed as a linear function of the state variables, i.e.,

$$P_t = \underbrace{\mathbf{p}^\top}_{2\times 5} \mathbf{s}_t + \underbrace{\bar{P}}_{2\times 1} \tag{8}$$

where the state variables are the position of intermediaries in each market, noise trader position in each market, and the fundamental value of the asset, i.e., $\mathbf{s}_t = (x_{C,t}, x_{S,t}, \beta_{C,t}, \beta_{S,t}, V_t)^\top$. Furthermore, we denote $\mathbf{p} = (\boldsymbol{\lambda_x}, \boldsymbol{\lambda_\beta}, \boldsymbol{\lambda}_V)$ where $\boldsymbol{\lambda_x}$ (2×2), $\boldsymbol{\lambda_\beta}$ (2×2) and $\boldsymbol{\lambda}_V$ (2×1) are the price loadings on the state variables respectively.

Let the state variable dynamics be

$$d\mathbf{s}_t = -\Gamma(\mathbf{s}_t - \bar{\mathbf{s}})dt + \Sigma^{\frac{1}{2}} d\mathbf{B}_t \tag{9}$$

where Σ is the instantaneous innovation covariance matrix, i.e., $\Sigma = \mathrm{Cov}(d\mathbf{s}_t)/dt$. The explicit expression of Σ is given by (54) in Appendix A.

Define

$$\boldsymbol{\Lambda} \equiv k(I - \boldsymbol{\zeta}\boldsymbol{\lambda_x}) \tag{10}$$

then the mean reversion matrix Γ is given by

$$\Gamma = \begin{pmatrix} \mathbf{\Lambda} & k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}}) - \boldsymbol{\eta}_\beta & -k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) \\ \mathbf{0}_{2\times 2} & \boldsymbol{\eta}_\beta & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 2} & \eta_v \end{pmatrix} \tag{11}$$

Throughout the analysis, we consider the equilibrium where $\mathbf{\Lambda}$ has positive eigenvalues, so that the intermediaries' holdings are stable.

The first order condition of the intermediaries' optimization problem is

$$\mathbb{E}_t \left[\begin{pmatrix} dP_{C,t} \\ dP_{S,t} \end{pmatrix} \right] = \mathbf{C} \begin{pmatrix} x_{C,t} \\ x_{S,t} \end{pmatrix} \tag{12}$$

where the marginal cost matrix $\mathbf{C}$ is given by

$$\mathbf{C} = \begin{pmatrix} \psi_C & 0 \\ 0 & \psi_S \end{pmatrix} + \gamma \underbrace{\mathbf{p}\Sigma\mathbf{p}^\top}_{=Var(d\mathbf{P}_t)} \tag{13}$$

Intuitively, the intermediary's expected return must equal the marginal cost of holding a given inventory position. The cost matrix $\mathbf{C}$ reflects two sources of friction: the risk-based cost, which scales with the intermediary's risk aversion and the variance of the portfolio, and the position cost, captured by the diagonal matrix with entries ψ_C and ψ_S, penalizing large asset positions regardless of portfolio composition.

Proposition 1 characterizes the equilibrium price in closed form. First of all, the price of each asset is increasing in the fundamental value V_t with a loading of 1. In addition, the constant terms of the prices $\bar{P}$ is determined by the natural demand of the institutional investors relative to the asset supply, scaled by the demand elasticities of the institutional investors. The price loading on the noise trader position $\boldsymbol{\lambda}_{\boldsymbol{\beta}}$, conditional on the intermediary sector's holdings, is given by the inverse of the demand elasticity matrix $\boldsymbol{\zeta}$. This is because fixing the intermediary position, the noise trader holding mirrors the holding of the institutional investors. Hence it's impact on the prices is determined by the demand elasticity of the institutional investors. Finally, the price loading on the intermediary position $\boldsymbol{\lambda}_{\boldsymbol{x}}$ is determined by the cost matrix $\mathbf{C}$ through a matrix equation. The price loading on the intermediary position captures how much the price needs to adjust to compensate the intermediaries for bearing extra inventory.

Proposition 1. *The equilibrium price is given by*

$$\begin{pmatrix} P_{C,t} \\ P_{S,t} \end{pmatrix} = \underbrace{\bar{\mathbf{P}}}_{\textit{Steady state}} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \underbrace{V_t}_{\textit{Fundamentals}} + \boldsymbol{\lambda_x} \cdot \underbrace{\begin{pmatrix} x_{C,t} \\ x_{S,t} \end{pmatrix}}_{\textit{Intermediary holdings}} + \boldsymbol{\lambda_\beta} \cdot \underbrace{\begin{pmatrix} \beta_{C,t} \\ \beta_{S,t} \end{pmatrix}}_{\textit{Noise trader holdings}} \tag{14}$$

where $\boldsymbol{\lambda_x}$ *is the solution to the following matrix equation*

$$-\boldsymbol{\lambda_x}\boldsymbol{\Lambda} = \begin{pmatrix} \psi_C & 0 \\ 0 & \psi_S \end{pmatrix} + \gamma \begin{pmatrix} \boldsymbol{\lambda_x} & \boldsymbol{\zeta}^{-1} & \mathbf{1} \end{pmatrix} \Sigma \begin{pmatrix} \boldsymbol{\lambda_x} \\ \boldsymbol{\zeta}^{-1} \\ \mathbf{1} \end{pmatrix} \tag{15}$$

$\boldsymbol{\lambda_\beta}$ *and* $\bar{\mathbf{P}}$ *are given by*

$$\boldsymbol{\lambda_\beta} = \boldsymbol{\zeta}^{-1} \tag{16}$$

$$\bar{\mathbf{P}} = \boldsymbol{\zeta}^{-1} \left(\boldsymbol{\theta} - \bar{\mathbf{S}} \right) \tag{17}$$

To emphasize the significance of intermediary frictions in creating asset price spreads, we examine a specific scenario where there is no position cost, i.e., $\psi_C = \psi_S = 0$. In this situation, there exists an equilibrium in which the two asset markets become perfectly integrated, resulting in identical prices for both assets. Consequently, the spread between these two assets is always zero. We characterize such equilibrium in Corollary 1. Our result directly supports the extensive empirical research that interprets relative spreads between two identical assets as indication of intermediary frictions. However, it also highlights that the spread is tied to a particular type of intermediary friction, namely the position cost, as the risk-based cost alone is not sufficient to generate a non-zero spread between the two assets.

Corollary 1. *When* $\psi_C = \psi_S = 0$*, there exists a solution where the two asset markets are perfectly integrated and the equilibrium price of the two assets are identical. The equilibrium price is given by*

$$P_{C,t} = P_{S,t} = \bar{P} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} V_t + \lambda_{\mathbf{x}} \cdot \begin{pmatrix} x_{C,t} \\ x_{S,t} \end{pmatrix} + \lambda_{\boldsymbol{\beta}} \cdot \begin{pmatrix} \beta_{C,t} \\ \beta_{S,t} \end{pmatrix} \tag{18}$$

where $\lambda_{\mathbf{x}}$ *is defined by* (70) *in Appendix* A *and* $\lambda_{\boldsymbol{\beta}}$ *is given by*

$$\lambda_{\boldsymbol{\beta}} = \frac{\lambda_{\mathbf{x}}}{\lambda_{\mathbf{x}} \boldsymbol{\zeta} (1,1)^\top} \tag{19}$$

Hence the equilibrium spread between the two assets is always zero.

As the two markets become perfectly integrated, the segmentation of preferred-habitat investors becomes irrelevant. This is evident from the term $\boldsymbol{\zeta}(1,1)^\top$, which represents the aggregate demand elasticity of institutional investors. In a perfectly integrated market, the demand elasticity only matters in aggregate, instead of its distribution across the two markets. In practice, we expect the position cost to be non-zero; therefore, we will consider the general case where $\psi_C > 0$ and $\psi_S > 0$ for the remainder of our analysis.

Next, we analyze the dynamics of price impact upon demand shocks. Proposition 2 characterizes the price and position dynamics following a demand shock, which are determined by the eigenvalues of the matrix $\boldsymbol{\Lambda}$, as defined in (10). To ensure stability, we need $\boldsymbol{\Lambda}$ to have positive eigenvalues. The magnitudes of the eigenvalues control the speed of convergence to long-run behavior.

Proposition 2. *The price and position dynamics following a demand shock is given by*

$$\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = -e^{-\boldsymbol{\Lambda}\tau} \tag{20}$$

$$\frac{\partial \mathbb{E}_t[\mathbf{P}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = -\boldsymbol{\lambda}_{\boldsymbol{x}} e^{-\boldsymbol{\Lambda}\tau} + \boldsymbol{\zeta}^{-1} \tag{21}$$

where

$$e^{-\boldsymbol{\Lambda}\tau} = \frac{e^{-\nu_1 \tau}}{\nu_2 - \nu_1}(\nu_2 \mathbf{I} - \boldsymbol{\Lambda}) + \frac{e^{-\nu_2 \tau}}{\nu_1 - \nu_2}(\nu_1 \mathbf{I} - \boldsymbol{\Lambda}), \tag{22}$$

and ν_1 *and* ν_2 *are the two eigenvalues of* $\boldsymbol{\Lambda}$.

Note that, on impact, the intermediary aborbs the entire shock, hence the intermediary position immediately after the shock is given by

$$\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top}\Big|_{\tau=0} = -\mathbf{I} \tag{23}$$

In the long run, the intermediary off loads the entire demand shock to the institutional

investors. Hence, the intermediary position reverts back to its steady state

$$\lim_{\tau\to\infty} \frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = \mathbf{0} \tag{24}$$

In terms of price dynamics, in the long run, the price response is only determined by the demand elasticity of the institutional investors, reflecting the fact that intermediaries are short-term investors in the market. That is,

$$\lim_{\tau\to\infty} \frac{\partial \mathbb{E}_t[\mathbf{P}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = \boldsymbol{\zeta}^{-1}. \tag{25}$$

That said, along the transition path, the price response does depend on the intermediary position. In fact, we can rewrite the price impact as

$$\mathbb{E}\left[\frac{\partial \boldsymbol{P}_{t+\tau}}{\partial \boldsymbol{\beta_t}^\top}\right] = \overbrace{\underbrace{\boldsymbol{\zeta}^{-1}}_{\text{Institution elasticity}}}^{\text{Long-run impact}} + \overbrace{\underbrace{-\partial \boldsymbol{x}_{t+\tau}/\partial \boldsymbol{\beta}_t^\top}_{\text{Intermediation share}} \times \underbrace{\boldsymbol{\lambda}_x}_{\text{Intermediary required comp.}}}^{\text{Transitory impact driven inventory dynamics}} \tag{26}$$

The price response reflects both a long-run component, which is determined by the demand elasticity of the preferred-habitat investors, and a transitory component, which is driven by the inventory dynamics of the intermediary sector. The transitory component captures the fact that intermediaries are short-term investors in the market, and they require compensation for bearing inventory in the short run.

The decomposition in equation (26) highlights why we cannot learn about the intermediary cost parameters just from the price impact from an identification perspective. The price impact depends on both the intermediary cost parameters as well as the demand elasticity of the institutional investors. We cannot separately identify them without additional moments. However, as we show in the next section, the time slope of the price impact only depends on the intermediary cost parameters, and hence becomes the crucial object for identification.

2.3 Volatility and Intermediary Costs

To explore the impact of various intermediary frictions on asset prices and their relative spreads, we analyze how the intermediary cost parameters $\boldsymbol{\psi}$ and γ influence the volatility of asset prices. To obtain explicit analytical solution, we focus on a special case where the two asset markets are symmetric, i.e., $\psi_C = \psi_S = \psi$, $\boldsymbol{\zeta}_{11} = \boldsymbol{\zeta}_{22}$, and the demand and fundamental

shocks are independent of each other. When the two markets are symmetric, we can define

$$\begin{pmatrix} M_l & 0 \\ 0 & M_s \end{pmatrix} \equiv \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \zeta^{-1} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\top} \tag{27}$$

where M_l and M_s are the long-run price impacts for the average price level and the relative spread, respectively.

We define $P_{l,t} = \frac{P_{C,t}+P_{S,t}}{2}$ as the average price of the two markets, and $P_{s,t} = \frac{P_{S,t}-P_{C,t}}{2}$ as the spread between the two markets. Furthermore, define V_l and V_s as the instantaneous volatility of $P_{l,t}$ and $P_{s,t}$, respectively.

$$V_l \equiv Var(dP_{l,t}) = \frac{1}{4}Var(dP_{C,t} + dP_{S,t}), \quad V_s \equiv Var(dP_{s,t}) = \frac{1}{4}Var(dP_{S,t} - dP_{C,t}) \tag{28}$$

We derive the following comparative statics results in Proposition 3. Both risk-based cost and gross position cost increase the volatility of the average price and the volatility of the spread. Intuitively, higher intermediation costs make it more costly for intermediaries to absorb shocks, hence they demand higher compensation, leading to more volatile price movements.

Proposition 3. *The instantaneous volatility of the average price V_l and the instantaneous volatility of the spread V_s are increasing in both the risk-aversion parameter γ and the position cost ψ, i.e.,*

$$\frac{\partial V_l}{\partial \gamma} > 0, \quad \frac{\partial V_l}{\partial \psi} > 0, \quad \frac{\partial V_s}{\partial \gamma} > 0, \quad \frac{\partial V_s}{\partial \psi} > 0 \tag{29}$$

However, even though both types of costs increase the volatility of price and spread, the relative magnitude is different. In Proposition 4, we show that the gross position cost increases the volatility of spread more relative to the volatility of the price level, while the risk-based cost matters more for the volatility of the price level than that of the spread.

Proposition 4. *The ratio between the instantaneous volatility of the spread and the instantaneous volatility of the average price, i.e., V_s/V_l, is increasing in the position cost ψ, i.e.,*

$$\frac{\partial (V_s/V_l)}{\partial \psi} > 0 \tag{30}$$

Furthermore, when $\frac{2M_s}{1-\gamma\sigma_\beta^2 M_s/k} < \frac{M_l}{1-\gamma\sigma_\beta^2 M_l/k}$, the ratio of the instantaneous volatility of the spread and the instantaneous volatility of the average price is decreasing in the risk-aversion

parameter γ, i.e.,

$$\frac{\partial(V_s/V_l)}{\partial\gamma} < 0 \tag{31}$$

Intuitively, to arbitrage the spread away between the two assets, the intermediaries need to take a hedged position $x_{C,t} - x_{S,t}$. This position nets out most of the risk from price movements but still requires the intermediaries to hold a large gross position. Hence the position cost has a large impact on this trading strategy, and therefore affects the spread directly. In contrast, to arbitrage the price level (when it deviates from the fundamental value), the intermediaries need to take a directional position in the two assets and bear risks from price movements. As a result, such trading strategy is more affected by the risk-based costs, and hence the risk aversion coefficient has a larger impact on the level of the prices.

An equivalent way to state the above result is through the instantaneous return correlation between the two assets, since the volatility ratio maps one-to-one into the correlation.

Corollary 2. *The instantaneous return correlation $\rho \equiv \text{Corr}(dP_{C,t}, dP_{S,t})$ is decreasing in the position cost ψ and, under the same sufficient condition as in Proposition 4, increasing in the risk-aversion parameter γ:*

$$\frac{\partial\rho}{\partial\psi} < 0, \qquad \frac{\partial\rho}{\partial\gamma} > 0 \tag{32}$$

Intuitively, risk-based costs *integrate* the two markets: they encourage arbitrageurs to take offsetting positions that hedge risk, linking the two markets through active arbitrage and raising return correlation. Gross position costs, by contrast, *segment* the markets by penalizing the gross positions required to do so, driving prices apart.

Our results in Proposition 4 and Corollary 2 highlight the significance of distinguishing between the two broad types of intermediary costs. While gross position costs have a substantial impact on the spread, risk-based costs are more influential for the price level. As risk-based costs rise, the wedge between spread volatility and price volatility widens, implying that the spread can miss much of the movements in the price level. By contrast, when position costs dominate, spread volatility tracks price volatility more closely, making the spread a more informative proxy for price-level effects. These findings make it essential to identify the two cost components separately. In the next section, we will discuss methods for empirically distinguishing between them using the price impact dynamics that follow a demand shock.

3 Identification and Estimation

We develop a new method for identifying intermediary costs from observed price dynamics following demand shocks. We show that the intermediaries' first-order condition links the initial rate of price reversion to marginal intermediation costs, providing a direct way to identify risk-based and gross position costs from observed price dynamics. We formalize this identification strategy below and then take it to the U.S. Treasury cash and OIS markets, using high-frequency demand shocks from Treasury auction result releases.

3.1 Identification from the Slope of Price Impact

As shown in (26) in the previous section, the price impact of demand shocks consists of two components: the long-run impact, determined by preferred-habitat investors' elasticity, and a transitory component, determined by the intermediary's inventory dynamics. As a result, we cannot identify the intermediary cost parameters ψ and γ directly from the price impact because it is a function of both the intermediaries' cost parameters and the institutional investors' demand elasticities. However, we below show that the cost parameters can be identified from the slope of the price impact curve following a demand shock.

Formally, the slope of the price impact curve is given by

$$\mathbb{E}\left[\frac{\partial}{\partial\tau}\left(\frac{\partial \boldsymbol{P}_{t+\tau}}{\partial \boldsymbol{\beta}_t^\top}\right)\right] = \underbrace{-\partial \boldsymbol{x}_{t+\tau}/\partial \boldsymbol{\beta}_t^\top}_{\text{Intermediation share}} \times \underbrace{\mathbf{C}}_{\text{marginal required comp.}} \tag{33}$$

and is determined by the intermediary's cost parameters alone. Intuitively, the slope of the price impact curve is the return that intermediaries get from bearing the inventory at a given instant, which should equal to their inventory share times the marginal cost of bearing the inventory. The logic of this identification strategy follows directly from the intermediary's first order condition.

With detailed quantity data on intermediary holdings, we can observe $-\partial \boldsymbol{x}_{t+\tau}/\partial \boldsymbol{\beta}_t^\top$ along the whole transition path and back out the cost parameters from the changes in the price impact of the cash and synthetic assets. Absent of such quantity data, we can still identify the cost parameter by leveraging the fact that the institutional investors have slow moving capital and the entire demand shock is absorbed by the intermediary on impact, i.e.,

$$\lim_{t\to 0}\mathbb{E}\left[\frac{\partial}{\partial\tau}\left(\frac{\partial \boldsymbol{P}_{t+\tau}}{\partial \boldsymbol{\beta}_t^\top}\right)\right] = \underbrace{-\partial \boldsymbol{x}_{t+\tau}/\partial \boldsymbol{\beta}_t^\top}_{\text{Intermediation share=100\%}} \times \underbrace{\mathbf{C}}_{\text{marginal required comp.}} = -\mathbf{C} \tag{34}$$

This means that even without detailed holdings data, we can still identify the cost parameters from the initial rate at which prices revert.

Figure 1 illustrates the intuition behind this identification strategy. The average price impact of a demand shock decomposes into two regions. The long-run level reflects institutional investors' demand elasticity, since they eventually absorb the entire demand shock. The additional transitory component above the long-run level reflects intermediary compensation: intermediaries absorb the shock on impact and require return compensation to hold the extra inventory while institutional investors gradually step in. The price impact therefore starts high on impact and decays toward the long-run level as intermediaries offload their positions. At each instant, the slope of the price impact curve equals the intermediaries' expected return on their inventory—their marginal revenue from providing liquidity. Under competitive intermediation, marginal revenue equals marginal cost in equilibrium, so the slope equals the intermediation share times the marginal cost of intermediation $\mathbf{C}$. On impact at $\tau = 0$, the intermediation share is 100%, so the initial slope directly reveals the marginal cost of intermediation.

This gives us a direct way to identify the intermediary cost matrix C without estimating the full economy. The identification comes directly from the intermediary's first order condition and the main assumption we need is that the entire demand shock is absorbed by the intermediary on impact. This identification approach can be applied to a wide range of models where the intermediary sector is the marginal supplier of liquidity.

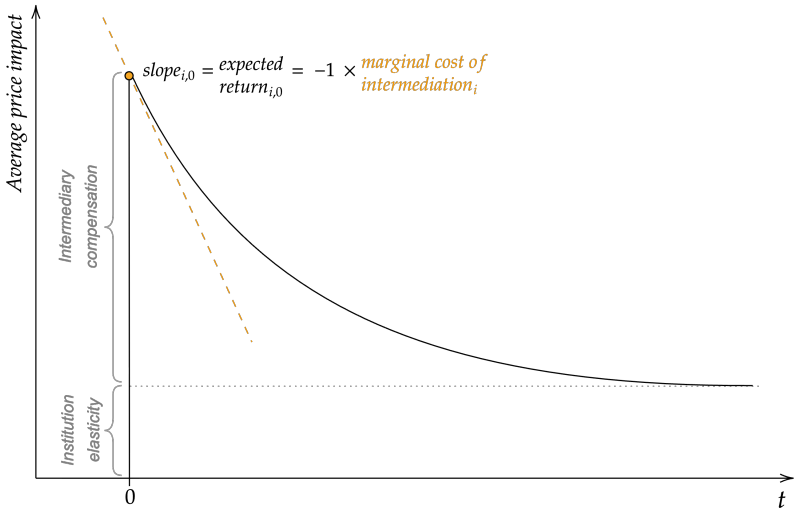


Figure 1: Identification from the initial decay rate. The figure illustrates the price impact dynamics following a demand shock. Initially, the entire shock is absorbed by intermediaries, so the intermediation share is 100%. The initial slope of the price impact curve therefore directly reveals the marginal cost matrix $\mathbf{C}$.

In our specific model, the intermediary cost matrix C has a particular structure, i.e.,

$$C = \begin{pmatrix} \psi + \gamma\sigma_C^2 & \gamma\sigma_{CS} \\ \gamma\sigma_{CS} & \psi + \gamma\sigma_S^2 \end{pmatrix} \tag{35}$$

where σ_C, σ_S, and σ_{CS} are all empirical observables. This allows us to back out the two cost parameters ψ and γ from the reversion rate of the price and spread response to demand shocks. Similar to the intuition on relative volatility ratio, the hedged position reflects the gross position cost more than risk-based costs, hence the reversion rate of the spread in response to the demand shock is more informative about ψ. On the other hand, the market position is more informative about the risk-based cost and hence the level of price response is more informative about γ. Combining the two allows to identify the two types of costs separately.

3.2 Treasury Yields and OIS Rates

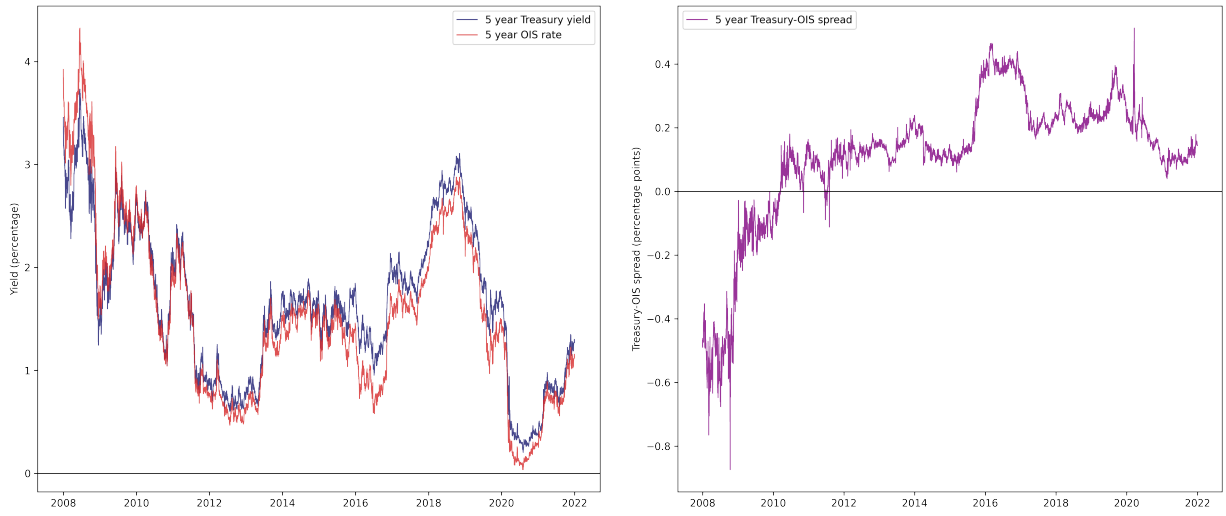
We measure the price level response using the 5-year constant-maturity Treasury yield, sourced from the U.S. Department of the Treasury's daily yield curve data. For the synthetic price level, we use the 5-year overnight index swap (OIS) rate from Bloomberg, where the underlying reference rate is the effective federal funds rate. The Treasury-OIS spread—the difference between the Treasury yield and the OIS rate—serves as our measure of the no-arbitrage spread.[4] The OIS rate provides a particularly clean synthetic price level benchmark because OIS contracts are derivatives settled in cash and do not require dealers to hold physical inventory on their balance sheets.

[4]The Treasury-OIS spread partly reflects the "inconvenience yield" of intermediating Treasuries: He, Nagel, and Song (2022) show that this spread captures the balance sheet costs borne by dealer-intermediaries when absorbing demand and supply shocks in the Treasury market.

Figure 2: 5-year Treasury yield and Treasury-OIS spread

(a) Treasury yield and OIS rate

(b) Treasury–OIS spread



The figure plots the 5-year constant-maturity Treasury yield, 5-year OIS rate, and the 5-year Treasury-OIS spread over the sample period. Panel (a) shows the Treasury yield and the OIS rate. Panel (b) shows the Treasury-OIS spread. The Treasury yield is from the U.S. Department of the Treasury's daily yield curve data. The OIS rate underlying the spread is the 5-year overnight index swap rate from Bloomberg, referenced to the effective federal funds rate.

Table 1: Summary statistics: Treasury yield and OIS rate

	Mean	Std. Dev.	Min	Median	Max	N
Treasury yield	1.66	0.74	0.20	1.64	3.73	3,621
OIS rate	1.54	0.84	0.04	1.47	4.33	3,621
Treasury-OIS spread	0.11	0.21	-0.87	0.14	0.51	3,621

Summary statistics for the 5-year constant-maturity Treasury yield, the 5-year OIS rate, and the Treasury-OIS spread. The sample spans January 2008 to January 2022. Treasury yield and OIS rate are in percent; the Treasury-OIS spread is in percentage points. The correlation between the Treasury yield and the OIS rate is 0.97.

Figure 2 plots the two series and the resulting spread over the sample period, and Table 1 reports summary statistics. The 5-year Treasury yield averages 1.66% over the sample, closely tracking the OIS rate (mean 1.54%), with a correlation of 0.97. The Treasury-OIS spread averages 11 basis points but exhibits substantial variation, ranging from −87 to 51

basis points, with notable spikes during the Global Financial Crisis and the COVID-19 episode visible in Panel (b).

3.3 Demand Shock Construction

We identify demand shocks to the Treasury market using the high-frequency price impact of Treasury auction result releases, following Droste, Gorodnichenko, and Ray (2024). In U.S. Treasury auctions, the total quantity to be issued is announced days in advance, but the composition of investor demand is not revealed until the auction closes and results are released. The yield change in a narrow window around the release of auction results therefore reflects unexpected shifts in investor demand, providing a measure of exogenous demand shocks to the Treasury market.

Figure 3 illustrates the construction. Let $y_{t,\text{pre}}$ denote the yield of the auctioned security measured at 12:50pm, approximately 10 minutes before the competitive bidding closes, and $y_{t,\text{post}}$ the yield at 1:10pm, after the results have been released. For reopenings, yields are measured on the outstanding (on-the-run) security; for new issues, yields are measured on the when-issued forward contract.[5] The demand shock on auction date t is defined as

$$\beta_{\text{Treasury},t} \propto u_{\text{Treasury},t} = y_{t,\text{post}} - y_{t,\text{pre}}. \tag{36}$$

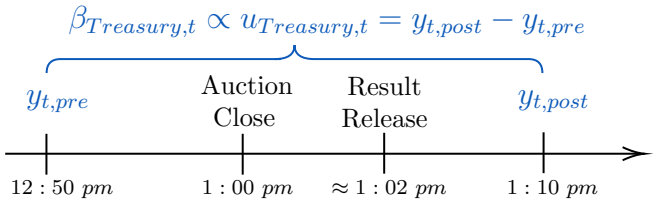
The intraday yield data are constructed from tick-level quotes sourced from GovPX, which provides real-time prices for on-the-run and when-issued Treasury securities. On non-auction days, the demand shock is set to zero.

An important caveat is that we do not directly observe the underlying demand shock $\beta_{\text{Treasury},t}$, but rather its price impact $u_{\text{Treasury},t} = y_{t,\text{post}} - y_{t,\text{pre}}$. As we discuss below, assuming the price impact is proportional to the demand shock, we can use it to identify the relative importance of risk-based versus gross position costs.

Conversely, weaker-than-expected demand raises yields. Figure 4 plots the demand shocks over the sample period, and Table 2 summarizes their distribution. The 165 shocks are approximately mean-zero (0.17 bps) with a standard deviation of 1.77 basis points, ranging from -5.55 to 8.50 basis points. The largest shocks are during the Global Financial Crisis and the Covid period, consistent with heightened uncertainty during these episodes. Note that a positive demand shock—more demand for Treasuries than expected—pushes bond prices up and hence yields down, so that $u_{\text{Treasury},t} < 0$.

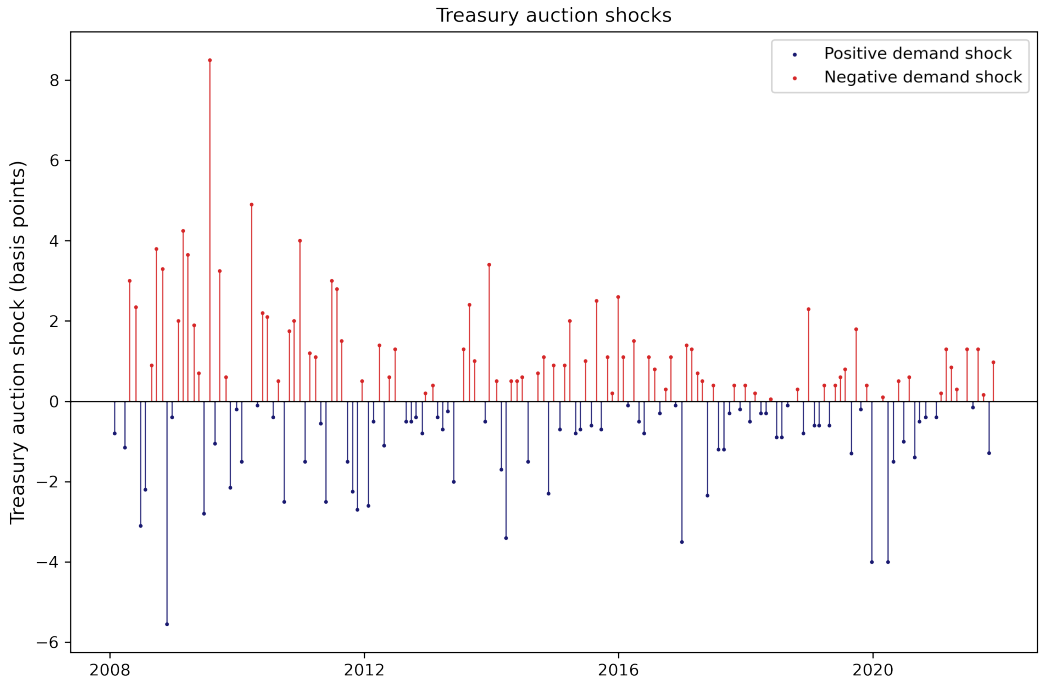
[5] When-issued contracts are forward contracts that settle on the issue date, providing a pre-issuance price for securities that do not yet trade in the secondary market.

Figure 3: Timeline of Treasury auction shock construction



The figure illustrates the construction of Treasury auction demand shocks following Droste, Gorodnichenko, and Ray (2024). The pre-auction yield $y_{t,\text{pre}}$ is measured at 12:50pm, before the competitive bidding closes at 1:00pm. The post-auction yield $y_{t,\text{post}}$ is measured at 1:10pm, after the results are released at approximately 1:02pm. The demand shock is $\beta_{\text{Treasury},t} \propto u_{\text{Treasury},t} = y_{t,\text{post}} - y_{t,\text{pre}}$.

Figure 4: Treasury auction demand shocks



The figure plots Treasury auction demand shocks for 5-year Treasury note auctions over the sample period (2008–2022). Each point represents the yield change $u_{\text{Treasury},t} = y_{t,\text{post}} - y_{t,\text{pre}}$ around a single auction result release, measured in basis points.

Table 2: Summary statistics: Treasury auction demand shocks

	Mean	Std. Dev.	Min	Median	Max	N
5Y auction shock (bps)	0.17	1.77	-5.55	0.05	8.50	165

Summary statistics for 5-year Treasury note auction demand shocks. The shock is the yield change from 12:50pm to 1:10pm around the release of auction results, measured in basis points. The sample spans January 2008 to January 2022.

3.3.1 Demand Shock and Identification

The identifying assumption underlying our demand shock measure is that the yield change around auction result releases reflects shifts in investor demand for Treasury securities that: (i) are exogenous to macroeconomic fundamentals, and (ii) hit the Treasury market but not the OIS market. The high-frequency nature of our identification makes both assumptions particularly plausible.

First, the revelation of fundamental information is unlikely to contaminate our shock measure. The narrow window of a few minutes around the auction result release rules out the arrival of macroeconomic news during the event window. A subtler issue is that auction participants may possess private information about fundamentals, so that their demand reveals fundamentals rather than reflecting an exogenous shift. This channel also appears unlikely: Droste, Gorodnichenko, and Ray (2024) show that surprise movements in auction demand are driven by relatively unsophisticated institutional investors—foreign monetary authorities, investment funds, and insurance companies—rather than hedge funds or broker-dealers who are more likely to possess private information.

Second, the high-frequency window implies that the demand shocks we capture originate from investors who face frictions in accessing the secondary Treasury market or OIS markets. If these investors could trade freely, they would have already expressed their demand in the secondary market or through OIS positions outside the auction window. The fact that their demand is revealed only at auction indicates they are restricted to the primary market, making it unlikely that they simultaneously generate demand shocks in OIS.

3.4 Mapping Yield and Spread Decay Rates to Intermediation Costs

We can test our model predictions using any two out of the three series—Treasury yields, OIS rates, or the Treasury-OIS spread—since any two can be used to construct the third. We focus on the yield level and the Treasury-OIS spread. Two considerations motivate this choice. First, the academic literature and policy makers frequently focus on the Treasury-OIS spread as an indicator of intermediation capacity in the Treasury market (e.g., He, Nagel, and Song, 2022), facilitating comparison with existing work. Second, our theory suggests that spreads provide a more direct measure of capacity costs, helping with statistical inference.

As derived in Section 3.1, intermediaries' first-order condition implies that the rate at which expected price levels revert following a demand shock is proportional to the cost matrix $\mathbf{C}$. We restate the first-order condition in terms of yields and spreads below.

Let $y_{1,t}$ denote the Treasury yield and $y_{2,t}$ the OIS rate. Denote the *instantaneous decay rates* of the yield level and the Treasury-OIS spread as

$$m_l := \frac{\partial}{\partial \tau} \frac{\partial \mathbb{E}_t[y_{1,t+\tau}]}{\partial \beta_{1,t}}\bigg|_{\tau=0}, \quad m_s := \frac{\partial}{\partial \tau} \frac{\partial \mathbb{E}_t[y_{1,t+\tau} - y_{2,t+\tau}]}{\partial \beta_{1,t}}\bigg|_{\tau=0}, \tag{37}$$

where $\beta_{1,t}$ is the demand shock to the Treasury market. Intuitively, m_l measures how fast the yield impact reverts (a more negative m_l indicates faster mean reversion), while m_s measures the same for the no-arbitrage spread.

From the first-order condition, these observable decay rates satisfy

$$\begin{pmatrix} m_l \\ m_s \end{pmatrix} = -\tilde{\lambda} \begin{pmatrix} \psi + \gamma\sigma_1^2 \\ \psi + \gamma\sigma_{1,s} \end{pmatrix}, \tag{38}$$

where $\sigma_1^2 := \mathrm{Var}(dy_{1,t})/dt$ is the instantaneous variance of Treasury yields, $\sigma_{1,s} := \mathrm{Cov}(dy_{1,t}, d(y_{1,t} - y_{2,t}))/dt$ is the covariance of yields and spreads, and $\tilde{\lambda} > 0$ is a positive scaling constant reflecting the fact that we don't observe the size of the demand shocks, only its price impact, which we assume to be proportional.[6]

Once these decay rates are estimated, we can quantify the relative importance of the risk and gross position costs. The two-equation system (38) can be inverted to express the structural parameters (up to the common scale $\tilde{\lambda}$) as functions of the observables m_l, m_s, and the yield covariance structure. Hence we can construct summary statistics that

[6] In addition, to map from prices to yields we use a first order approximation around the bonds steady state duration D and price level P which are also absorbed in $\tilde{\lambda}$.

measure the fraction of marginal intermediation cost attributable to risk-based costs. The risk contribution to the intermediation costs associated with arbitraging the Treasury yield and the Treasury-OIS spread are,

$$\hat{C}_{r,l} \equiv \frac{\gamma\sigma_1^2}{\psi + \gamma\sigma_1^2} = \frac{\sigma_1^2}{\sigma_1^2 - \sigma_{1,s}} \left(1 - \frac{m_s}{m_l}\right), \tag{39}$$

$$\hat{C}_{r,s} \equiv \frac{\gamma\sigma_{1,s}}{\psi + \gamma\sigma_{1,s}} = \frac{\sigma_{1,s}}{\sigma_1^2 - \sigma_{1,s}} \left(\frac{m_l}{m_s} - 1\right). \tag{40}$$

The contribution of the gross position cost is one minus the contribution of the risk-based cost.[7]

To build intuition, consider the following two cases. If Treasury yields and OIS rates move in lockstep after a demand shock—so that the spread does not respond ($m_s = 0$)—then the entire yield reversion is attributable to risk costs ($\hat{C}_{r,l} = 100\%$). This is because arbitraging the spread is nearly risk-free, so if the spread shows no movement, it must be that risk costs, not gross position costs, are dominating intermediation. Conversely, if the spread also reverts ($m_s \neq 0$), then gross position costs contribute to the yield response, and $\hat{C}_{r,l} < 100\%$. Hence, testing whether $m_s = 0$ is a sufficient condition to see if gross position costs matter.

3.5 Estimation Procedure

Local projections. We estimate the impulse response of yields and spreads to Treasury auction demand shocks using the local projection framework of Jordà (2005). For each outcome variable $i \in \{l, s\}$ (yield level and Treasury-OIS spread, respectively) and horizon $\tau = 0, 1, \ldots, H$, we estimate

$$\Delta_\tau y_{i,t} = a_i + \theta_{i,\tau}\, u_{i,t} + c_{i,\tau}\, \Delta_\tau y_{i,t-\tau-1} + \varepsilon_{i,t+\tau}, \tag{41}$$

where $\Delta_\tau y_{i,t} := y_{i,t+\tau} - y_{i,t}$ is the cumulative change from day t to day $t+\tau$ and $u_{i,t}$ is the demand shock.[8] The coefficient $\theta_{i,\tau}$ traces out the impulse response function (IRF) at horizon τ.

[7] Because the risk-based cost contributions depend only on the ratio m_s/m_l and the yield covariance structure, they are identified without knowledge of the demand shock magnitude; the scale of the shock cancels in the ratio.

[8] The lagged dependent variable $\Delta_\tau y_{i,t-\tau-1}$ addresses autocorrelation in the errors, following **olea2021local**.

Polynomial decay approximation. The local projection coefficients $\theta_{i,\tau}$ are estimated separately for each horizon and are therefore noisy. To recover the instantaneous decay rate $m_i = \partial\theta_i(\tau)/\partial\tau|_{\tau=0}$ reliably, we smooth the IRF by imposing a polynomial structure. Specifically, we parameterize $\theta_i(\tau) = \sum_{k=0}^{K} b_{i,k}\,\tau^k$ and estimate the polynomial coefficients jointly via the panel regression

$$\Delta_\tau y_{i,t} = a_i + \sum_{k=0}^{K} b_{i,k}\,(\tau^k\,u_{i,t}) + c_{i,\tau}\,\Delta_\tau y_{i,t-\tau-1} + \varepsilon_{i,t+\tau}, \quad \tau = 0,\ldots,H;\ i = l, s. \tag{42}$$

The instantaneous decay rate is then simply the linear coefficient,

$$\hat{m}_i = \frac{\partial\hat{\theta}_i(\tau)}{\partial\tau}\Big|_{\tau=0} = \sum_{k=0}^{K} k\cdot\hat{b}_{i,k}\,\tau^{k-1}\Big|_{\tau=0} = \hat{b}_{i,1}$$

We estimate the level and spread equations jointly, which directly provides the covariance between $\hat{m}_l$ and $\hat{m}_s$ needed for inference on the risk contributions.[9] Standard errors for the decay rates $\hat{m}_l$ and $\hat{m}_s$ are computed using date-clustered covariance matrices, which account for the correlation across horizons within each auction date.

Mapping decay rates to contribution. Contribution of risk to yields and spreads must lie between 0% and 100%. We impose these bounds when estimating the risk contributions using constrained optimization and use a date-cluster bootstrap for inference; the details are described in Appendix B.3.[10]

Support identification. In practice, the IRF may not begin decaying immediately—for example, if slow-moving investors wait a few days before they begin to enter the market—and may reach its long-run level before the end of the estimation horizon. We use a data-driven procedure to identify the relevant estimation window $[\underline{\tau}, \overline{\tau}]$: we fit a high-order polynomial over the full IRF horizon, locate peaks and troughs in the IRF via the polynomial's derivatives, and select the window that captures the decay region. See detailed algorithm description and robustness checks in Appendix B.2.

[9] The baseline specification uses a linear decay function ($K = 1$). We show robustness to quadratic ($K = 2$) specifications and to a theory-implied exponential decay approach in Table 3; see Appendix B.1 for details on the exponential decay estimation.

[10] We report the unconstrained decay rate estimates and the constrained contribution estimates. Overall, we find constraints have negligible impact on local projection fit and slope estimates, but mainly address issue of near zero decay rates inducing noise in contribution estimation. See Appendix B.3 for details.

4 Estimation Results

We present the main findings in this section. Section 4.1 presents the baseline results for the full sample from 2008 to 2022. The findings highlight that although yields move considerably in response to Treasury demand shocks, no-arbitrage spreads show negligible movement, indicating that on average risk-based costs are the primary constraint on intermediation capacity. Section 4.2 then zooms into two crisis episodes to explore whether the costs at play differ during periods of market stress. For the 2008 crisis, we find that no-arbitrage spreads do not respond, and hence risk-based costs remain the main factor inhibiting intermediation capacity—consistent with the regulatory environment not yet imposing costs on gross position sizes. During the COVID crisis, we find some evidence that no-arbitrage spreads respond, suggesting that gross position costs also play a role. Our findings suggest that while risk-based costs generally dominate intermediation capacity, post-GFC regulations imposing costs on gross position sizes may become an additional constraint during periods of market stress.

4.1 Baseline Results

Figure 5 shows the impulse response functions to Treasury auction demand shocks for the full sample (2008–2022). Panel (a) shows that Treasury yields and the OIS rate move in lockstep: both initially increase and then slowly decay toward zero. This is consistent with slow-moving investors entering the market over time and hence increasing its absorptive capacity. In line with the lockstep movement, panel (b) shows that the Treasury-OIS spread does not respond to the shock. This pattern indicates that risk-based costs, rather than gross position costs, are the main driver of intermediation capacity.

The intuition is as follows: arbitraging the Treasury-OIS spread is nearly risk-free, since going long one asset and short the other hedges away risk. Risk-based costs therefore do not inhibit this arbitrage trade. Gross position costs, by contrast, penalize long/short positions regardless of risk exposure. The absence of a spread response thus indicates active arbitrage and suggests that gross position costs play a limited role on average—that is, in normal times, which constitute most of the sample. We turn next to quantitatively confirming this interpretation and then show that it does not hold during periods of recent market stress.

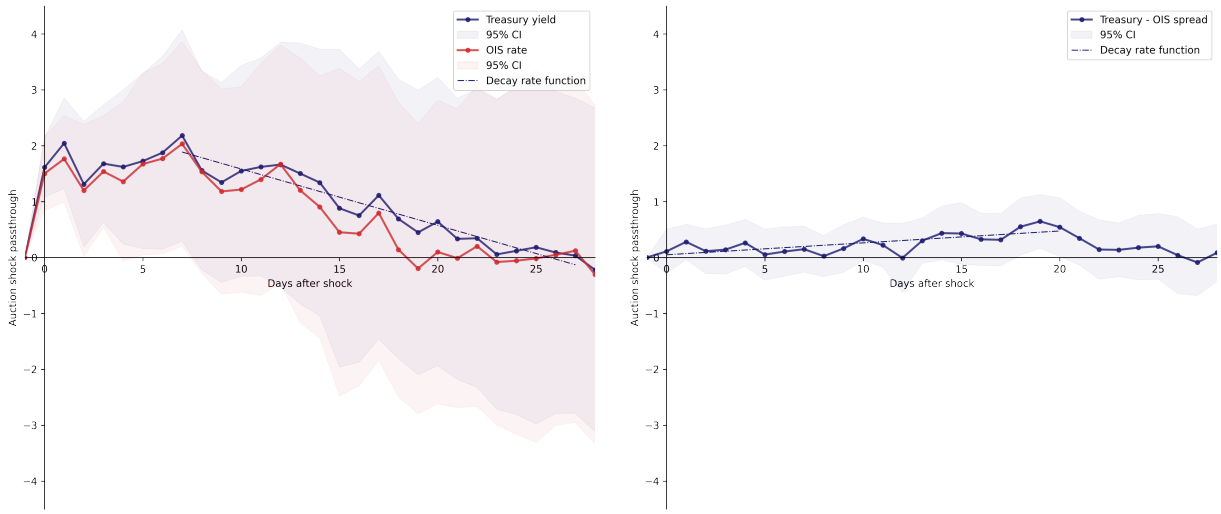
Table 3 presents the risk contribution estimates for the 5-year note auction demand shock across several specifications. The risk contribution measures the share of the yield's instantaneous decay rate attributable to risk-based costs, as defined in Section 3.1.[11]

[11]The estimation window for the decay function is automatically detected using a polynomial support

Figure 5: Impulse response to Treasury auction demand shocks: Full sample (2008–2022)

(a) Treasury yield and OIS rate

(b) Treasury–OIS spread



The figures plot impulse response functions of 5-year Treasury yields, the OIS rate, and the Treasury–OIS spread to Treasury auction demand shocks over the full sample (2008–2022). The demand shock is constructed from 5-year Treasury note auctions. Panel (a) shows the response of Treasury yields (blue) and the OIS rate (red), with 95% confidence bands. The dashed line shows the fitted decay function. Panel (b) shows the response of the Treasury–OIS spread. The estimation window for the decay function is automatically detected using a polynomial support identification approach. The sample spans $N = 3{,}621$ trading days with 156 auction dates.

The baseline specification uses a linear decay function estimated over a 20-day window. The 95% confidence interval for the risk contribution is 95% to 100%, meaning that essentially the entire yield response can be attributed to risk-based costs. This is consistent with the visual evidence in Figure 5: the yield and OIS rate move in lockstep while the spread remains flat, indicating that the spread's decay rate is negligible relative to the yield's decay rate.

These findings are robust across alternative specifications. Shortening the estimation window to 15 or 10 days yields risk contribution confidence intervals of 92%–100% and 78%–100%, respectively. The confidence interval excludes zero across all linear specifications (indicated in bold in Table 3), confirming that risk-based costs account for the majority of the yield response. Allowing for a quadratic decay function produces a wider confidence interval of 36%–100%, though the point estimate remains consistent with risk-based costs dominating.

identification approach described in Section 3.5. The shorter window specifications use the auto-detected start point but manually adjust the end point.

Table 3: Estimation results: Full sample (2008–2022)

	Risk contribution (95% CI)	Function	Window
Baseline	**95% – 100%**	Linear	20 days
Window (15d)	**92% – 100%**	Linear	15 days
Window (10d)	**78% – 100%**	Linear	10 days
Quadratic	**36% – 100%**	Quadratic	20 days

The table reports 95% confidence intervals for the risk contribution of the yield response to Treasury auction demand shocks across different specifications. The risk contribution measures the share of the yield's instantaneous decay rate attributable to risk-based costs. The sample spans $N = 3{,}621$ trading days from January 2008 to January 2022, with 156 auction dates. The demand shock is constructed from 5-year Treasury note auctions. Standard errors use date-cluster bootstrap (1,000 replications). Risk contribution confidence intervals use the delta method, bounded to $[0, 100]$. Bold confidence intervals indicate that the lower bound exceeds zero, providing statistically informative evidence that risk-based costs account for a positive share of the yield response.

4.2 The Story of Two Crises

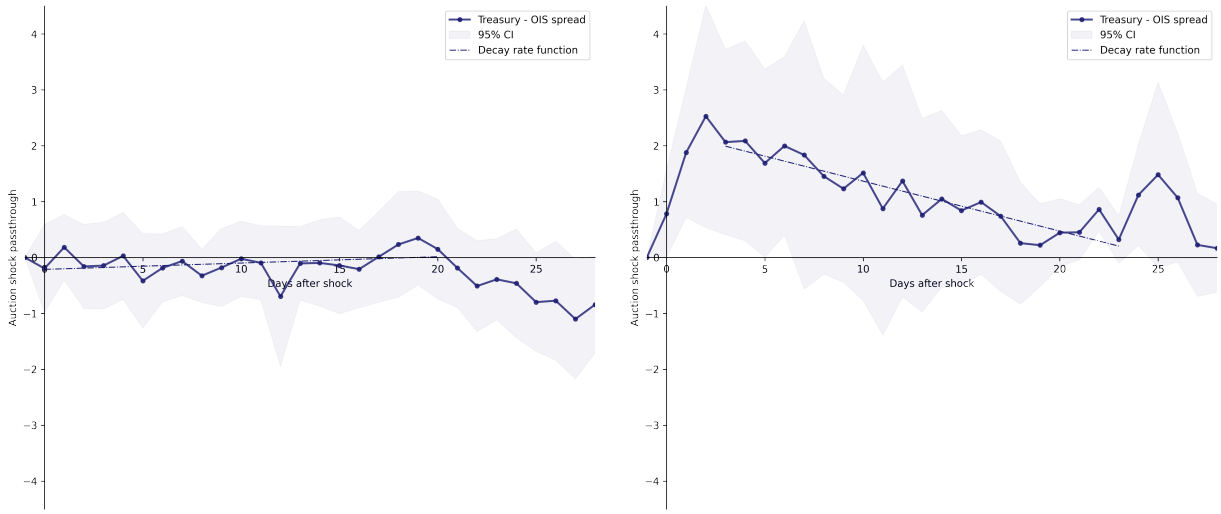
We now examine two crisis episodes—the Global Financial Crisis (GFC, 2008–2010) and the COVID-19 crisis (2020)—to test whether the dominant cost inhibiting intermediation has changed following the introduction of gross position regulatory constraints. In 2017, the Supplementary Leverage Ratio, enacted as part of the Basel III reforms, began imposing regulatory costs on broker-dealers' gross positions. The key diagnostic is the spread response: if the Treasury-OIS spread does not respond, risk-based costs remain dominant; if it responds, gross position costs have become the dominant constraint. Notably, preliminary evidence suggests that even in the post-Basel III normal-times sample (2017–2019), the spread does not respond, indicating that gross position constraints bind only during periods of market stress rather than in normal times.

Figure 6 shows the spread impulse responses for the two episodes. Panel (a) shows the GFC episode: the Treasury-OIS spread shows no systematic response to auction demand shocks, remaining close to zero throughout the 30-day horizon. This mirrors the full-sample result and indicates that risk-based costs remain the binding constraint on intermediation capacity during the GFC. Panel (b) presents a different picture for the COVID crisis: the spread shows a notable response on impact and decays over the following weeks. This suggests that gross position costs also play a role in constraining intermediation capacity

during the COVID episode.

Figure 6: Spread impulse response to Treasury auction demand shocks: Crisis episodes

(a) GFC (Jan 2008 to Jan 2010) (b) COVID (Feb 2020 to Feb 2021)



The figures plot impulse response functions of the Treasury–OIS spread to 5-year Treasury note auction demand shocks for two crisis episodes. Panel (a) shows the GFC episode (January 2008 to January 2010). Panel (b) shows the COVID episode (February 2020 to February 2021). Both panels show 95% confidence bands and the fitted decay function. The corresponding impulse responses for Treasury yields and OIS rates are reported in Appendix C.

The economic interpretation is as follows. During the GFC, dealer balance sheets were under severe stress, but the regulatory environment did not yet impose explicit gross position constraints (the Supplementary Leverage Ratio and other post-crisis regulations had not yet been implemented). Risk-based costs—capital adequacy, Value-at-Risk limits—were the dominant constraints, consistent with the spread showing no response. The COVID crisis, by contrast, occurred in a post-Dodd-Frank regulatory environment where the Supplementary Leverage Ratio and other balance sheet constraints impose costs on gross position sizes regardless of risk. When dealer balance sheets came under stress in March 2020, these gross position constraints may have become binding, contributing to the spread response.

5 Calibration

To quantify the effects of risk-based and gross position costs on market outcomes, we complement the reduced-form estimation in Section 4 with a structural calibration. Using the

calibrated model, we conduct two counterfactual exercises. The first applies the model to the COVID-19 episode, where our reduced-form results indicate that gross position costs were the dominant driver of intermediation capacity. The second examines the broader implications of relaxing different regulatory tools—such as the SLR versus the Stress Capital Buffer—for Treasury market functioning.

5.1 Calibration Moments

We calibrate the model using unconditional moments in the two markets instead of the conditional moments from demand shocks. While the IRFs from demand shocks are informative about cost ratios and are robust to model mis-specifications, they are too noisy to identify the complete set of parameter precisely. Hence we leverage the unconditional moments, in particular, the variance profiles of the yield and the spread, for identifying the structural parameters.

The variance profiles of the yield level and the Treasury-OIS spread are defined as the per-period variance of H-day changes as a function of the horizon H. Using the definition of price level and spread in Section 2, our model implies that these profiles take the parametric form

$$V_i(H) = V_i^\infty + (V_i - V_i^\infty)\frac{1 - e^{-\nu_i H}}{\nu_i H}, \quad i \in \{l, s\}, \tag{43}$$

where V_i is the short-run (instantaneous) variance, V_i^∞ is the long-run variance, and ν_i governs how quickly the variances decrease by horizon. More specifically, the long-run variance of the price level V_l^∞ is driven by the fundamental shock variance σ_V^2 and the permanent demand shock variance σ_β^2, while the long-run spread variance V_s^∞ is driven solely by the permanent demand shock variance σ_β^2.

$$V_l^\infty = \tfrac{1}{2}\sigma_\beta^2 M_l^2 + \sigma_V^2, \tag{44}$$

$$V_s^\infty = \tfrac{1}{2}\sigma_\beta^2 M_s^2. \tag{45}$$

The instantaneous variance V_i is amplified relatively to the long-run variance as the intermediary's inventory dynamics create transitory price pressure in response to demand shocks.

$$V_l = \frac{1}{2}\sigma_\beta^2 B_l^2 + \sigma_V^2 \tag{46}$$

$$V_s = \frac{1}{2}\sigma_\beta^2 B_s^2. \tag{47}$$

The degree of amplification B_l and B_s depends on the structural parameters, with higher costs and slower institutional adjustment leading to greater amplification. The exact expressions for B_l and B_s are derived in Appendix A.4. Futhermore, the mean-reversion rate ν_l and ν_s corresponds to the two eigenvalues of $\mathbf{\Lambda}$ defined in Proposition 2.

We jointly identify the structural parameters $(\gamma, \psi, k, \sigma_\beta, \sigma_V, M_s)$ from the empirical moments of the variance profiles. We externally calibrate M_l from Chaudhary, Fu, and Zhou, 2025, who estimate the long-run multiplier of the Treasury market using GIV methods.[12]

Estimation of the Variance Profile. We estimate the variance profile in three steps. First, we compute the empirical per-period variance $\widehat{V}_i(H) = \text{Var}(\Delta_H y_i)/H$ at each horizon using overlapping multi-day differences. For the yield level, we residualize the 5-year Treasury yield on the 3-month Treasury yield before computing variances; this removes the persistent fundamental component from the yield.[13] This ensures the estimation method is valid even if the fundamentals do not follow a martingale.[14] Second, we fit (43) to the empirical estimates via nonlinear least squares, recovering the six profile parameters $(V_l, V_l^\infty, \nu_l, V_s, V_s^\infty, \nu_s)$. Third, we convert all variance moments to price-return units using the estimated modified duration.

Figure 7 displays the fitted variance profiles. Both level and spread profiles fit the exponential decay function well. The spread profile (right panel) exhibits substantially more mean reversion than the level profile (left panel): the ratio $V_s/V_s^\infty \approx 6.4$ compared with $V_l/V_l^\infty \approx 1.5$. The spread is a hedged position with low fundamental risk, so demand shocks create large transitory spread distortions that only diminish once institutional capital has adjusted. Table 4 reports the estimated variance profile parameters.

Calibrated parameters. Table 5 reports the calibrated structural parameters. We find that the gross position cost is small relative to the risk cost ($\psi = 0.03$ vs. $\gamma = 0.87$), consistent with the estimation result from Section 4 that risk contributions account for approximately 95% of yield reversion over the full sample. Furthermore, the spread long-run multiplier $M_s = 0.06$ is much smaller than the level multiplier $M_l = 0.15$, reflecting the substitution effect in long-term investors' demand.

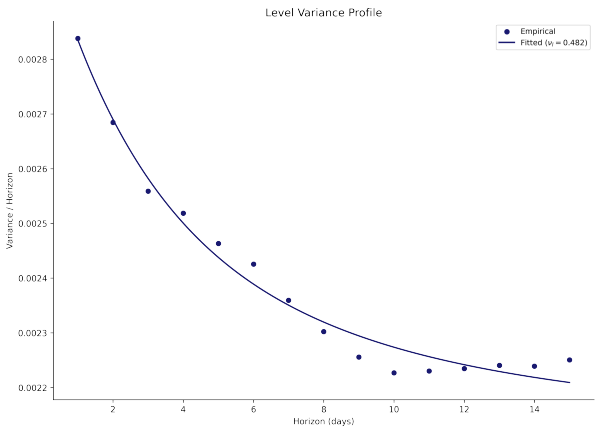
[12] The choice of M_l is innocuous as the system is invariant to the rescaling of M_l.

[13] We use the fixed-beta method: regress daily yield changes on daily 3-month yield changes over the full sample and subtract the fitted values.

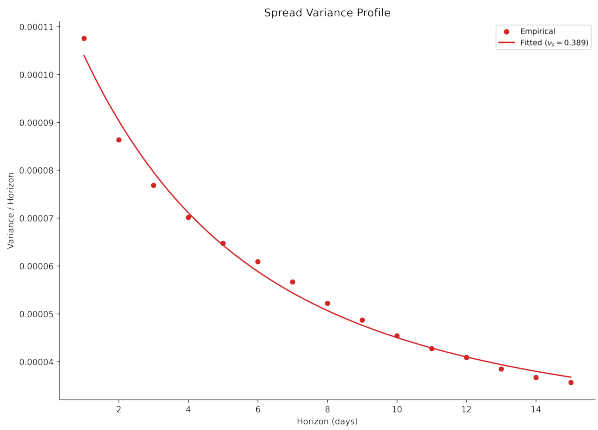
[14] Our results are robust to residualizing spreads as well using the same method.

Figure 7: Variance profiles of Treasury yields and Treasury-OIS spread

(a) Yield level



(b) Treasury-OIS spread



Each panel plots the empirical per-period variance ratio $\widehat{V}_i(H) = \mathrm{Var}(\Delta_H y_i)/H$ (dots) and the fitted parametric profile (43) (curve) as a function of horizon H in days. The yield level is the 5-year constant-maturity Treasury yield residualized on the 3-month yield. The sample spans January 2008 to January 2022.

Table 4: Estimated variance profile moments

Moment	Description	Value
V_s	Short-run spread variance	1.2
V_s^∞	Long-run spread variance	0.19
ν_s	Spread variance decay rate	0.39
V_l	Short-run level variance	30
V_l^∞	Long-run level variance	21
ν_l	Level variance decay rate	0.48

Estimated parameters of the variance profile (43) for the yield level ($i = l$) and Treasury-OIS spread ($i = s$). V_i is the short-run instantaneous variance, V_i^∞ is the long-run permanent component, and ν_i is the mean-reversion rate. Variance units are in squared basis points per day (yield space); decay rates are in day^{-1}. The sample spans January 2008 to January 2022.

Table 5: Calibrated structural parameters

Parameter	Description	Value
γ	Risk cost	0.87
ψ	Gross position cost	0.03
k	Slow-moving capital speed	0.16
σ_β	Demand shock volatility	0.47
σ_V	Fundamental shock volatility	0.21
M_s	Spread demand multiplier	0.06
M_l	Level demand multiplier (external)	0.15

Calibrated structural parameters recovered from matching the estimated variance profile moments in Table 4. The long-run level multiplier $M_l = 0.15$ is set externally following Chaudhary, Fu, and Zhou (2025).

5.2 Counterfactual Analysis

With the structural parameters in hand, we conduct two counterfactual experiments that illustrate how changes in the two cost components propagate asymmetrically into yield volatility and spread volatility, as well as their responses to demand shocks. The first quantifies the COVID-19 episode, where our reduced-form results indicate that gross position costs were the dominant friction. The second examines the broader implications of regulatory changes for Treasury market functioning.

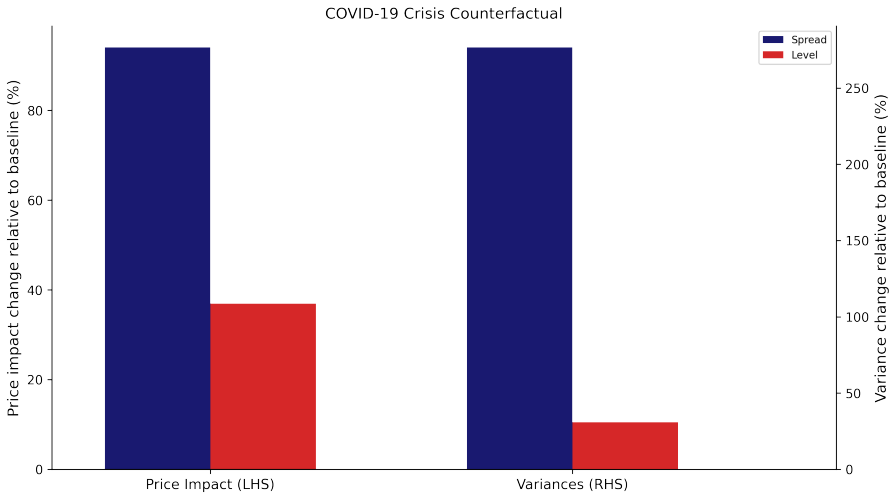
5.2.1 COVID-19 Crisis

Our first counterfactual zooms in on the COVID-19 episode of 2020, during which Treasury market functioning deteriorated sharply and intermediary balance sheets came under severe pressure. The stress was alleviated quickly due to the Federal Reserve's aggressive interventions. We are interested in learning about the market outcomes if the gross position costs had remained elevated for longer periods of time. To this end, we adjust the gross position cost of the model to match the empirically observed IRF slope ratio m_s/m_l during the crisis window, and we compute the implied effects on price impact and short-run variances due to this change in the gross position cost.

Our estimation result in Section 4 suggests that the gross-position cost relative to the risk cost was much higher during the COVID crisis than over the full sample, reflecting the severe balance-sheet pressure dealers faced during the dash-for-cash episode in March 2020. As we increase the gross position cost to match the COVID-period slope ratio, our model

implies a much larger increase in the spread response to demand shocks than the yield response. Figure 8 displays the results. While the elevated gross position cost increases yield price impact by only 37%, it increases spread price impact to demand shocks by more than 90%. This is consistent with the empirical observation that the Treasury-OIS spread widened dramatically in March 2020, as the selling pressure built up in the Treasury market. The changes in short-run variance are even more dramatic: the elevated gross position cost increases spread volatility by more than 250%, while increasing yield volatility by only 30%. These results reflect the fact that the gross position cost has a disproportionate effect on spread relative to yield. So while spreads directionally reflected increased distortions in the Treasury yield during COVID-19, they overstated the magnitude of yield distortions.

Figure 8: COVID-19 counterfactual: implied effects of elevated gross position costs



The figure displays counterfactual price impact changes and short-run variance changes for the yield level and Treasury-OIS spread under a COVID-consistent parameterization. The gross position cost ψ is recalibrated to match the COVID-period IRF slope ratio m_s/m_l; all other parameters are held at their full-sample calibrated values from Table 5.

5.2.2 Banking Deregulation

We now ask how Treasury market outcomes would change under a relaxation of intermediary constraints. Specifically, we compute model-implied outcomes under two scenarios: a 50% reduction in ψ, the gross position cost (as would arise from relaxing the SLR), holding γ fixed; and a 50% reduction in γ, the risk cost (as would arise from reducing the Stress Capital Buffer), holding ψ fixed. For each counterfactual, we resolve the model to obtain

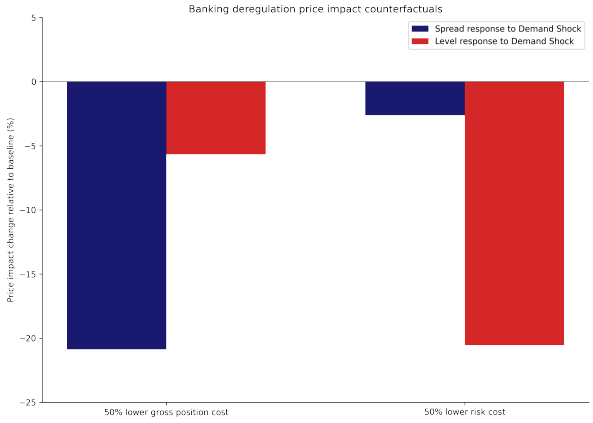
the implied short-run amplification factors due to intermediary frictions, and compute the resulting percentage changes in price impact of demand shocks and short-run variance.

Figure 9 displays the results. The left panel shows percentage changes in price impact, and the right panel shows percentage changes in short-run variance. A sharp asymmetry emerges: reducing the gross position cost ψ compresses the spread response to demand shocks substantially, by more than 20%, while having little effect on yield response — reducing it by only 5%. On the other hand, reducing the risk cost γ reduces yield price impact considerably by more than 20%, while leaving spread response almost unaffected. The changes in short-run volatility paint a consistent picture: reduction in ψ reduces spread volatility by more than 35%, yet has very little impact on the yield volatility. These findings reflect our theoretical findings in Proposition 4 that the gross position cost disproportionately affect spread volatility while the risk cost disproportionately affects yield volatility.

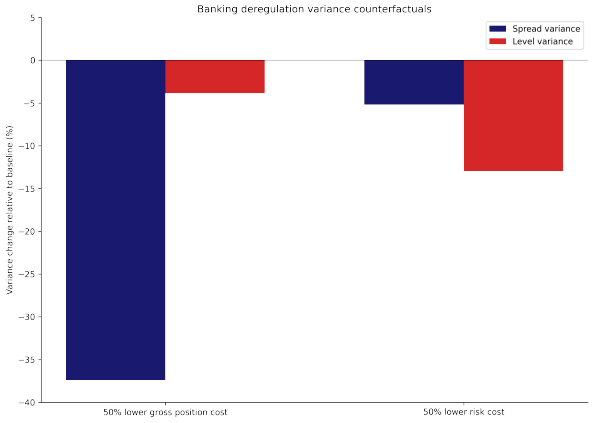
These results have direct regulatory implications. Relaxing gross position costs operates mainly through ψ and therefore has a large effect on the Treasury-OIS spread but a smaller effect on outright yield volatility and Treasury market liquidity. Relaxing risk-based constraints operates through γ and mainly affects yield levels.

Figure 9: Counterfactual effects of reducing intermediary costs by 50%

(a) Change in price impact



(b) Change in short-run variance



Each panel shows the percentage change in price impact amplification (left) and short-run variance (right) for the yield level and Treasury-OIS spread under two counterfactual scenarios: a 50% reduction in the gross position cost ψ and a 50% reduction in the risk cost γ. All other parameters are held at their calibrated values from Table 5. We resolve the model in each case and generate the implied effects on the Treasury markets.

6 Conclusion

Spreads between nearly identical assets are widely used as evidence of intermediary frictions and as signals of market dysfunction. This paper shows that spreads may not always be a useful diagnostic for price level distortions: when risk-based costs are the dominant friction on intermediation capacity, price levels move substantially while spreads remain quiescent. Recognizing this distinction requires jointly modeling price levels and spreads and separately identifying the underlying costs.

We develop a model to jointly analyze how asset prices and relative spreads are affected by two broad types of intermediation costs: risk-based costs tied to portfolio risk and gross position costs tied to the size of individual asset positions. The model delivers a sharp theoretical distinction between the two. Risk-based costs disproportionately affect the level of asset prices, because trading against price-level deviations requires intermediaries to take directional, unhedged positions that expose them to portfolio risk. Gross position costs, by contrast, have a larger effect on spreads, because arbitraging the spread requires hedged positions that are costly in terms of gross balance sheet usage but largely neutral in risk. Absent gross position costs, the model admits an equilibrium where two identical assets are perfectly integrated and the spread is always zero, highlighting the tight link between spreads and position costs.

Our sufficient statistic approach exploits the intermediaries' first-order condition: the initial decay rate of prices following an exogenous demand shock equals the marginal cost of intermediation. Because the two types of costs affect yields and spreads differently, combining the initial decay rates of yields and spreads separately identifies each cost component. This approach requires only that intermediaries are the marginal liquidity providers on impact—a common feature across intermediation models—and does not require detailed data on intermediary holdings.

We apply this framework to the U.S. Treasury cash and OIS markets, using high-frequency demand shocks from Treasury auction result releases. In the full sample (2008–2022), risk-based costs are the dominant driver of intermediation capacity. A diagnostic based solely on spreads would miss the risk-based frictions that primarily govern price-level dynamics in the Treasury market. These findings hold unconditionally, but the relative importance of the two costs shifts during crises. During the Global Financial Crisis, risk-based costs remain dominant, consistent with the pre-Basel III regulatory environment. During the COVID crisis, by contrast, gross position costs become the dominant driver of intermediation capacity, consistent with the post-2017 implementation of the Supplementary Leverage Ratio imposing balance sheet costs on dealer positions.

To quantify these findings structurally, we complement the reduced-form estimation with a calibration of the model to unconditional variance profiles of Treasury yields and the Treasury-OIS spread. The calibrated parameters confirm that risk-based costs dominate gross position costs in the full sample, consistent with our reduced-form evidence. We conduct two counterfactual exercises. First, we apply the calibrated model to the COVID-19 episode, adjusting the gross position cost to match the observed crisis-period cost ratio. We find that the elevated gross position cost almost doubled the spread's price impact and increased spread volatility by over 250%, compared with much more modest increases for yields. So while spreads directionally reflected increased distortions during COVID-19, they overstated the magnitude of yield distortions. Second, motivated by the current policy discussion of relaxing bank regulations, we examine the broader implications of reducing each cost component. Our results reveal a sharp asymmetry: a reduction in the gross position cost (as from relaxing the SLR) compresses the spread response to demand shocks and spread volatility substantially, while leaving yield dynamics nearly unchanged; a reduction in the risk cost (as from reducing the Stress Capital Buffer) does the opposite, materially reducing yield price impact with little effect on the spread. These counterfactuals provide direct structural evidence on how different regulatory tools affect Treasury market functioning differently.

Our findings carry direct regulatory implications. Policymakers who monitor spreads as the primary indicator of market dysfunction may underestimate the extent of illiquidity when risk-based costs dominate, as these costs manifest primarily in price levels rather than in spreads. More broadly, the informativeness of spreads as a diagnostic depends on the nature of the regulatory regime: regulations that limit portfolio risk lead to very different spread and price dynamics compared to regulations that restrict gross position sizes. Our framework provides a useful way to assess both dimensions of intermediation capacity and to evaluate how different regulatory tools affect each. Going forward, it would be valuable to extend this analysis to other asset classes and to incorporate time-varying costs, which would allow for a more comprehensive understanding of how intermediation frictions evolve across market segments and over time.

References

Adrian, Tobias, Erkko Etula, and Tyler Muir (Dec. 2014). "Financial Intermediaries and the Cross-Section of Asset Returns". In: *The Journal of Finance* 69.6, pp. 2557–2596. ISSN: 0022-1082, 1540-6261. DOI: 10.1111/jofi.12189.

Adrian, Tobias and Hyun Song Shin (Feb. 1, 2014). "Procyclical Leverage and Value-at-Risk". In: *The Review of Financial Studies* 27.2, pp. 373–403. ISSN: 0893-9454. DOI: 10.1093/rfs/hht068.

Bank of England (2025). *Enhancing the Resilience of the Gilt Repo Market*. Discussion Paper. Bank of England. URL: https://www.bankofengland.co.uk/paper/2025/discussion-paper/enhancing-the-resilience-of-the-gilt-repo-market.

Barth, Daniel and R. Jay Kahn (Oct. 1, 2025). "Hedge Funds and the Treasury Cash-Futures Basis Trade". In: *Journal of Monetary Economics* 155, p. 103823. ISSN: 0304-3932. DOI: 10.1016/j.jmoneco.2025.103823.

Board of Governors of the Federal Reserve System (May 15, 2020). *Financial Stability Report: May 2020*. Board of Governors of the Federal Reserve System.

Boyarchenko, Nina et al. (July 1, 2018). *Bank-Intermediated Arbitrage*. DOI: 10.2139/ssrn.3200041. Social Science Research Network: 3200041. Pre-published.

Brunnermeier, Markus K. and Yuliy Sannikov (Feb. 2014). "A Macroeconomic Model with a Financial Sector". In: *American Economic Review* 104.2, pp. 379–421. DOI: 10.1257/aer.104.2.379.

Chaudhary, Manav, Julie Zhiyu Fu, and Haonan Zhou (Oct. 2025). "Anatomy of the Treasury Market: Who Moves Yields?" Working paper. URL: https://fuzhiyu.me/TreasuryGIVPaper/Treasury_GIV_draft.pdf.

Droste, Matthew, Yuriy Gorodnichenko, and Walker D. Ray (2024). "Unbundling Quantitative Easing: Taking a Cue from Treasury Auctions".

Du, Wenxin, Benjamin Hébert, and Amy Wang Huber (Apr. 1, 2023). "Are Intermediary Constraints Priced?" In: *The Review of Financial Studies* 36.4, pp. 1464–1507. ISSN: 0893-9454. DOI: 10.1093/rfs/hhac050.

Du, Wenxin, Benjamin Hébert, and Wenhao Li (Dec. 1, 2023). "Intermediary Balance Sheets and the Treasury Yield Curve". In: *Journal of Financial Economics* 150.3, p. 103722. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2023.103722.

Du, Wenxin, Alexander Tepper, and Adrien Verdelhan (2018). "Deviations from Covered Interest Rate Parity". In: *The Journal of Finance* 73.3, pp. 915–957. ISSN: 1540-6261. DOI: 10.1111/jofi.12620.

Duffie, Darrell (2010). "Presidential Address: Asset Price Dynamics with Slow-Moving Capital". In: *The Journal of Finance* 65.4, pp. 1237–1267. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.2010.01569.x.

Duffie, Darrell et al. (Aug. 2023). *Dealer Capacity and U.S. Treasury Market Functionality*. Staff Reports (Federal Reserve Bank of New York). Federal Reserve Bank of New York. DOI: 10.59576/sr.1070.

Duffie, James Darrell (2023). "Resilience Redux in the U.S. Treasury Market". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.4552735.

Gabaix, Xavier and Ralph S. J. Koijen (June 28, 2021). *In Search of the Origins of Financial Fluctuations: The Inelastic Markets Hypothesis*. w28967. National Bureau of Economic Research. DOI: 10.3386/w28967.

Gârleanu, Nicolae and Lasse Heje Pedersen (June 1, 2011). "Margin-Based Asset Pricing and Deviations from the Law of One Price". In: *The Review of Financial Studies* 24.6, pp. 1980–2022. ISSN: 0893-9454. DOI: 10.1093/rfs/hhr027.

Greenwood, Robin, Samuel G Hanson, and Gordon Y Liao (Apr. 2018). "Asset Price Dynamics in Partially Segmented Markets". In: *The Review of Financial Studies* 31.9, pp. 3307–3343. ISSN: 0893-9454. DOI: 10.1093/rfs/hhy048. eprint: https://academic.oup.com/rfs/article-pdf/31/9/3307/25521746/hhy048.pdf.

Gromb, Denis and Dimitri Vayanos (Nov. 1, 2002). "Equilibrium and Welfare in Markets with Financially Constrained Arbitrageurs". In: *Journal of Financial Economics*. Limits on Arbitrage 66.2, pp. 361–407. ISSN: 0304-405X. DOI: 10.1016/S0304-405X(02)00228-3.

Haddad, Valentin and Tyler Muir (Dec. 2021). "Do Intermediaries Matter for Aggregate Asset Prices?" In: *The Journal of Finance* 76.6, pp. 2719–2761. ISSN: 0022-1082, 1540-6261. DOI: 10.1111/jofi.13086.

— (2025). "Intermediaries and Asset Prices". In.

Hanson, Samuel G., Aytek Malkhozov, and Gyuri Venter (Apr. 1, 2024). "Demand-and-Supply Imbalance Risk and Long-Term Swap Spreads". In: *Journal of Financial Economics* 154, p. 103814. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2024.103814.

He, Zhiguo, Bryan Kelly, and Asaf Manela (Oct. 1, 2017). "Intermediary Asset Pricing: New Evidence from Many Asset Classes". In: *Journal of Financial Economics* 126.1, pp. 1–35. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2017.08.002.

He, Zhiguo and Arvind Krishnamurthy (Apr. 2013). "Intermediary Asset Pricing". In: *American Economic Review* 103.2, pp. 732–70. DOI: 10.1257/aer.103.2.732.

He, Zhiguo and Arvind Krishnamurthy (Nov. 1, 2018). "Intermediary Asset Pricing and the Financial Crisis". In: *Annual Review of Financial Economics* 10 (Volume 10, 2018),

pp. 173–197. ISSN: 1941-1367, 1941-1375. DOI: 10.1146/annurev-financial-110217-022636.

He, Zhiguo, Stefan Nagel, and Zhaogang Song (Jan. 1, 2022). "Treasury Inconvenience Yields during the COVID-19 Crisis". In: *Journal of Financial Economics* 143.1, pp. 57–79. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2021.05.044.

Jordà, Òscar (2005). "Estimation and Inference of Impulse Responses by Local Projections". In: *American Economic Review* 95.1, pp. 161–182. ISSN: 0002-8282. DOI: 10.1257/0002828053828518.

Koijen, Ralph S. J. and Motohiro Yogo (2019). "A Demand System Approach to Asset Pricing". In: *Journal of Political Economy* 127.4, pp. 1475–1515. ISSN: 0022-3808. DOI: 10.1086/701683.

Santos, Tano and Pietro Veronesi (Aug. 1, 2022). "Leverage". In: *Journal of Financial Economics* 145 (2, Part B), pp. 362–386. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2021.09.001.

Vayanos, Dimitri and Jean-Luc Vila (2021). "A Preferred-Habitat Model of the Term Structure of Interest Rates". In: *Econometrica : journal of the Econometric Society* 89.1, pp. 77–112. DOI: 10.3982/ECTA17440. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA17440.

A Derivations and Proofs

Law of motion. From market clearing condition we have

$$\mathbf{x}_t + \mathbf{y}_t + \boldsymbol{\beta}_t = \bar{\mathbf{S}} \tag{48}$$

The law of motion for $dy_{i,t}$ is given by

$$dy_{i,t} = k(Z_{i,t} - y_{i,t})dt \tag{49}$$

Plug in the expression for $Z_{i,t}$ and write in vector form we have

$$d\mathbf{y}_t = k(-\boldsymbol{\zeta}\mathbf{P}_t + V_t\boldsymbol{\zeta}\mathbf{1} + \boldsymbol{\theta} - \mathbf{y}_t)dt \tag{50}$$

substituting the market clearing condition we have

$$d\mathbf{x}_t = -k\Big(\mathbf{x}_t - \boldsymbol{\zeta}\mathbf{P}_t + V_t\boldsymbol{\zeta}\mathbf{1} + \boldsymbol{\theta} - \bar{\mathbf{S}} + \boldsymbol{\beta}_t\Big)dt - d\boldsymbol{\beta}_t. \tag{51}$$

Next, substitute (8) in (72), we get,

$$d\mathbf{x}_t = -k\Big(\mathbf{x}_t - \boldsymbol{\zeta}\mathbf{p}\mathbf{s}_t - \boldsymbol{\zeta}\bar{\mathbf{P}} + V_t\boldsymbol{\zeta}\mathbf{1} + \boldsymbol{\theta} - \bar{\mathbf{S}} + \boldsymbol{\beta}_t\Big)dt - d\boldsymbol{\beta}_t. \tag{52}$$

$$= -k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{x}})\mathbf{x}_t dt - k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}})\boldsymbol{\beta}_t dt + k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1})V_t dt + \text{constant}dt - d\boldsymbol{\beta}_t. \tag{53}$$

This implies the instantaneous innovation covariance matrix is given by

$$\Sigma \equiv \frac{\text{Cov}(d\mathbf{s}_t)}{dt} = \begin{pmatrix} \sigma_C^2 & \sigma_{CS} & -\sigma_C^2 & -\sigma_{CS} & -\sigma_{CV} \\ \sigma_{CS} & \sigma_S^2 & -\sigma_{CS} & -\sigma_S^2 & -\sigma_{SV} \\ -\sigma_C^2 & -\sigma_{CS} & \sigma_C^2 & \sigma_{CS} & \sigma_{CV} \\ -\sigma_{CS} & -\sigma_S^2 & \sigma_{CS} & \sigma_S^2 & \sigma_{SV} \\ -\sigma_{CV} & -\sigma_{SV} & \sigma_{CV} & \sigma_{SV} & \sigma_V^2 \end{pmatrix} \tag{54}$$

and the mean reversion matrix Γ is given by

$$\Gamma = \begin{pmatrix} \boldsymbol{\Lambda} & k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}}) - \boldsymbol{\eta}_\beta & -k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) \\ \mathbf{0}_{2\times 2} & \boldsymbol{\eta}_\beta & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 2} & \eta_v \end{pmatrix} \tag{55}$$

where $\mathbf{\Lambda} = k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{x}})$, and $\boldsymbol{\eta}_\beta = \begin{pmatrix} \eta_{\beta_C} & 0 \\ 0 & \eta_{\beta_S} \end{pmatrix}$.

A.1 Proof of Proposition 1

To solve the equilibrium price, plug in the conjectured price function into the first order condition (12), we have

$$-\mathbf{p}\Gamma(\mathbf{s}_t - \bar{\mathbf{s}}) = \mathbf{C}\mathbf{x}_t \tag{56}$$

$$-\begin{pmatrix} \boldsymbol{\lambda}_{\boldsymbol{x}} & \boldsymbol{\lambda}_{\boldsymbol{\beta}} & \boldsymbol{\lambda}_V \end{pmatrix} \begin{pmatrix} \mathbf{\Lambda} & k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}}) & -k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) \\ \mathbf{0}_{2\times 2} & \boldsymbol{\eta}_\beta & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 2} & \eta_v \end{pmatrix} \begin{pmatrix} \mathbf{x}_t - \bar{\mathbf{x}} \\ \boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}} \\ V_t - \bar{V} \end{pmatrix} = \mathbf{C}\mathbf{x}_t \tag{57}$$

Expanding the equation we get

$$-\boldsymbol{\lambda}_{\boldsymbol{x}}\mathbf{\Lambda}(\mathbf{x}_t - \bar{\mathbf{x}}) - \boldsymbol{\lambda}_{\boldsymbol{\beta}}\boldsymbol{\eta}_\beta(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}) - \boldsymbol{\lambda}_V\eta_v(V_t - \bar{V}) - k\boldsymbol{\lambda}_{\boldsymbol{x}}(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}})(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}) \tag{58}$$

$$+ k\boldsymbol{\lambda}_{\boldsymbol{x}}\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1})(V_t - \bar{V}) = \mathbf{C}\mathbf{x}_t \tag{59}$$

Matching the coefficients we have

$$[\mathbf{x}_t]: \quad -\boldsymbol{\lambda}_{\boldsymbol{x}}\mathbf{\Lambda} = \mathbf{C} \tag{60}$$

$$[\boldsymbol{\beta}_t]: \quad -\boldsymbol{\lambda}_{\boldsymbol{\beta}}\boldsymbol{\eta}_\beta - k\boldsymbol{\lambda}_{\boldsymbol{x}}(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}}) = \mathbf{0} \tag{61}$$

$$[V_t]: \quad -\boldsymbol{\lambda}_V\eta_v + k\boldsymbol{\lambda}_{\boldsymbol{x}}\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) = \mathbf{0} \tag{62}$$

Consider the limiting case where $\eta_{\beta_i} \to 0$ and $\eta_v \to 0$. When $\boldsymbol{\lambda}_{\boldsymbol{x}}$ has full rank, from (61) and (62) we must have

$$\boldsymbol{\lambda}_{\boldsymbol{\beta}} = \boldsymbol{\zeta}^{-1} \tag{63}$$

$$\boldsymbol{\lambda}_V = \mathbf{1} \tag{64}$$

To solve for $\boldsymbol{\lambda}_{\boldsymbol{x}}$, we can the expression for $\boldsymbol{\lambda}_{\boldsymbol{\beta}}$ and $\boldsymbol{\lambda}_V$ back into $\mathbf{C}$ in (60)

$$-\boldsymbol{\lambda}_{\boldsymbol{x}}\mathbf{\Lambda} = \begin{pmatrix} \psi_C & 0 \\ 0 & \psi_S \end{pmatrix} + \gamma \begin{pmatrix} \boldsymbol{\lambda}_{\boldsymbol{x}} & \boldsymbol{\zeta}^{-1} & \mathbf{1} \end{pmatrix} \Sigma \begin{pmatrix} \boldsymbol{\lambda}_{\boldsymbol{x}} \\ \boldsymbol{\zeta}^{-1} \\ \mathbf{1} \end{pmatrix} \tag{65}$$

in which the only unknown is $\boldsymbol{\lambda}_{\boldsymbol{x}}$. This is a quadratic matrix equation in $\boldsymbol{\lambda}_{\boldsymbol{x}}$ and is only well-defined when $\psi_C > 0$ and $\psi_S > 0$.

Finally, matching the coefficients on the constant term we have $\bar{x} = 0$, which implies

$$\bar{\mathbf{P}} = \boldsymbol{\zeta}^{-1}\left(\boldsymbol{\theta} - \bar{\mathbf{S}}\right) \tag{66}$$

A.2 Proof of Corollary 1

When $\psi_C = \psi_S = 0$, the above solution does not work. We hypothesize that the two assets comove perfectly, and verify that this is indeed a solution.

First, when the two assets comove perfectly, $\mathbf{C} = c\mathbf{1}\mathbf{1}^\top$. Hence from (60), we have

$$\boldsymbol{\lambda}_{\boldsymbol{x}} = -c\mathbf{1}\mathbf{1}^\top\boldsymbol{\Lambda}^{-1} \tag{67}$$

$$(1,-1)^\top\boldsymbol{\lambda}_{\boldsymbol{x}} = 0 \quad \Rightarrow \quad \lambda_{C,\mathbf{x}} = \lambda_{S,\mathbf{x}} = \lambda_{\mathbf{x}} \tag{68}$$

Plug this into (61), we have

$$\boldsymbol{\lambda}_{\boldsymbol{\beta}} = \frac{\lambda_{\mathbf{x}}}{\lambda_{\mathbf{x}}\boldsymbol{\zeta}(1,1)^\top} \tag{69}$$

From (62) we still have $\boldsymbol{\lambda}_V = \mathbf{1}$. Plug the coefficients back into (60) and multiply both sides by $\begin{pmatrix}1 & 0\end{pmatrix}$ we get

$$-\lambda_{\mathbf{x}}k(I - \boldsymbol{\zeta}\begin{pmatrix}1\\1\end{pmatrix}\lambda_{\mathbf{x}}) = \gamma\begin{pmatrix}\lambda_{\mathbf{x}} & \frac{\lambda_{\mathbf{x}}}{\lambda_{\mathbf{x}}\boldsymbol{\zeta}(1,1)^\top} & 1\end{pmatrix}\Sigma\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}\lambda_{\mathbf{x}} & \frac{\lambda_{\mathbf{x}}}{\lambda_{\mathbf{x}}\boldsymbol{\zeta}(1,1)^\top} & 1\end{pmatrix} \tag{70}$$

This is a quadratic system in $\lambda_{\mathbf{x}}$ with two equations and two unknowns.

A.3 Proof of Proposition 2

In the case when $\boldsymbol{\psi} \neq \mathbf{0}$, we can plug in the solution for prices into the law of motion for $\mathbf{x}_t$ in (53) to get

$$d\mathbf{x}_t = -\boldsymbol{\Lambda}\mathbf{x}_t dt - d\beta_t + \text{constant}dt \tag{71}$$

For stationarity, we need the eigenvalues of $\boldsymbol{\Lambda}$ to be positive, which ensures $e^{-\boldsymbol{\Lambda}T} \to 0$ as $T \to \infty$. Hence the price impact following a demand shock decays exponentially at rate $\boldsymbol{\Lambda}$.

$$\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = -e^{-\boldsymbol{\Lambda}\tau} \tag{72}$$

where

$$e^{-\boldsymbol{\Lambda}\tau} = \frac{e^{-\nu_1\tau}}{\nu_2 - \nu_1}(\nu_2\mathbf{I} - \boldsymbol{\Lambda}) + \frac{e^{-\nu_2\tau}}{\nu_1 - \nu_2}(\nu_1\mathbf{I} - \boldsymbol{\Lambda}) \tag{73}$$

Here ν_1 and ν_2 are the two eigenvalues of $\boldsymbol{\Lambda}$, and they control the speed of convergence to long-run behavior.

Next, we analyze the price dynamics following a demand shock. Using

$$\mathbf{P}_t = \boldsymbol{\lambda}_{\boldsymbol{x}}\mathbf{x}_t + \boldsymbol{\lambda}_{\boldsymbol{\beta}}\boldsymbol{\beta}_t + \boldsymbol{\lambda}_V V_t + \bar{\mathbf{P}} \tag{74}$$

$$\frac{\partial \mathbb{E}_t[\mathbf{P}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = \boldsymbol{\lambda}_{\boldsymbol{x}}\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} + \boldsymbol{\lambda}_{\boldsymbol{\beta}} \tag{75}$$

Plug in $\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top}$ from (72) and $\boldsymbol{\lambda}_{\boldsymbol{\beta}}$ from (63), we have

$$\frac{\partial \mathbb{E}_t[\mathbf{P}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = -\boldsymbol{\lambda}_{\boldsymbol{x}}e^{-\boldsymbol{\Lambda}\tau} + \boldsymbol{\zeta}^{-1} \tag{76}$$

A.4 Proof of Proposition 3

Given the definition of $P_{l,t}$ and $P_{s,t}$,

$$Var(dP_{l,t})/dt = \frac{1}{4}(1,1)\Sigma_P \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{4}(1,1)\mathbf{p}^\top\Sigma\mathbf{p} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{77}$$

$$Var(dP_{s,t})/dt = \frac{1}{4}(-1,1)\Sigma_P \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \frac{1}{4}(-1,1)\mathbf{p}^\top\Sigma\mathbf{p} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \tag{78}$$

Since the two markets are symmetric, we denote $\sigma_C = \sigma_S = \sigma_\beta$. Given the shock correlations are zero, we have

$$\Sigma = \begin{pmatrix} \sigma_\beta^2 & 0 & -\sigma_\beta^2 & 0 & 0 \\ 0 & \sigma_\beta^2 & 0 & -\sigma_\beta^2 & 0 \\ -\sigma_\beta^2 & 0 & \sigma_\beta^2 & 0 & 0 \\ 0 & -\sigma_\beta^2 & 0 & \sigma_\beta^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_V^2 \end{pmatrix} \tag{79}$$

For convenience, we define $E = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$. Given the symmetry of the two markets,

$$\lambda_{C,x_C} = \lambda_{S,x_S} \quad \lambda_{C,x_S} = \lambda_{S,x_C} \tag{80}$$

$$\lambda_{C,\beta_C} = \lambda_{S,\beta_S} \quad \lambda_{C,\beta_S} = \lambda_{S,\beta_C} \tag{81}$$

$$\lambda_{C,V} = \lambda_{S,V} \tag{82}$$

Hence $E^\top \boldsymbol{\lambda_x} E$ and $E^\top \boldsymbol{\lambda_\beta} E$ are diagonal matrices, we define

$$\begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix} \equiv E^\top \boldsymbol{\lambda_x} E, \qquad \begin{pmatrix} M_l & 0 \\ 0 & M_s \end{pmatrix} \equiv E^\top \boldsymbol{\lambda_\beta} E \tag{83}$$

Recall that Proposition 1 gives $\boldsymbol{\lambda_\beta} = \boldsymbol{\zeta}^{-1}$ and symmetry implies that $\boldsymbol{\zeta}^{-1}$ is diagonalized by the level/spread basis E, the entries M_l and M_s are exactly the level and spread eigenvalues of $\boldsymbol{\lambda_\beta}$, i.e., the long-run price impacts of the level and spread demand shocks. Hence we can write

$$d \begin{pmatrix} P_{l,t} \\ P_{s,t} \end{pmatrix} = \frac{1}{\sqrt{2}} E^\top \mathbf{p}(\mathbf{s}) = \frac{1}{2} \begin{pmatrix} l_x & l_x \\ -s_x & s_x \end{pmatrix} d\mathbf{x}_t + \frac{1}{2} \begin{pmatrix} M_l & M_l \\ -M_s & M_s \end{pmatrix} d\boldsymbol{\beta}_t + \frac{1}{2} \begin{pmatrix} 2 \\ 0 \end{pmatrix} dV_t \tag{84}$$

$$Var(d \begin{pmatrix} P_{l,t} \\ P_{s,t} \end{pmatrix})/dt = \frac{1}{4} \begin{pmatrix} l_x & l_x & M_l & M_l & 2 \\ -s_x & s_x & -M_s & M_s & 0 \end{pmatrix} \Sigma \begin{pmatrix} l_x & -s_x \\ l_x & s_x \\ M_l & -M_s \\ M_l & M_s \\ 2 & 0 \end{pmatrix} \tag{85}$$

To expand, first compute Σ times each column of the right matrix. For column 1:

$$\Sigma \begin{pmatrix} l_x \\ l_x \\ M_l \\ M_l \\ 2 \end{pmatrix} = \sigma_\beta^2 (l_x - M_l) \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 2\sigma_V^2 \end{pmatrix} \tag{86}$$

Dotting with each row of the left matrix:

$$\begin{aligned}
\text{Row 1:}\quad & 2l_x\sigma_\beta^2(l_x - M_l) - 2M_l\sigma_\beta^2(l_x - M_l) + 4\sigma_V^2 = 2\sigma_\beta^2(l_x - M_l)^2 + 4\sigma_V^2 \\
\text{Row 2:}\quad & s_x\sigma_\beta^2(l_x - M_l) - s_x\sigma_\beta^2(l_x - M_l) - M_s\sigma_\beta^2(l_x - M_l) + M_s\sigma_\beta^2(l_x - M_l) = 0
\end{aligned}$$

For column 2:

$$\Sigma \begin{pmatrix} -s_x \\ s_x \\ -M_s \\ M_s \\ 0 \end{pmatrix} = \sigma_\beta^2(s_x - M_s) \begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \\ 0 \end{pmatrix} \tag{87}$$

Dotting with each row:

$$\begin{aligned}
\text{Row 1:}\quad & -l_x\sigma_\beta^2(s_x - M_s) + l_x\sigma_\beta^2(s_x - M_s) + M_l\sigma_\beta^2(s_x - M_s) - M_l\sigma_\beta^2(s_x - M_s) = 0 \\
\text{Row 2:}\quad & 2s_x\sigma_\beta^2(s_x - M_s) - 2M_s\sigma_\beta^2(s_x - M_s) = 2\sigma_\beta^2(s_x - M_s)^2
\end{aligned}$$

Combining with the $\frac{1}{4}$ prefactor, the covariance matrix is diagonal:

$$\text{Var}\left(d\begin{pmatrix} P_{l,t} \\ P_{s,t} \end{pmatrix}\right)/dt = \frac{1}{2}\begin{pmatrix} \sigma_\beta^2(l_x - M_l)^2 + 2\sigma_V^2 & 0 \\ 0 & \sigma_\beta^2(s_x - M_s)^2 \end{pmatrix} \tag{88}$$

Level and spread returns are instantaneously uncorrelated. The variances are

$$\text{Var}(dP_{l,t})/dt = \frac{1}{2}\left(\sigma_\beta^2(l_x - M_l)^2 + 2\sigma_V^2\right) \tag{89}$$

$$\text{Var}(dP_{s,t})/dt = \frac{1}{2}\sigma_\beta^2(s_x - M_s)^2 \tag{90}$$

These variance formulas correspond to the average-price and half-spread objects $(P_{l,t}, P_{s,t})$. To solve for (l_x, s_x), however, it is more convenient to rotate (60) by the orthonormal basis E itself. That rotated fixed-point system is written for the level and spread eigen-components, so no additional $\frac{1}{2}$ enters the Riccati equations below. To solve for l_x and s_x, we multiply

both sides of (60) by E

$$-\begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix} k(I - \tilde{\boldsymbol{\zeta}} \begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix}) = \begin{pmatrix} \psi & 0 \\ 0 & \psi \end{pmatrix} + \gamma \begin{pmatrix} l_x & l_x & M_l & M_l & 2 \\ -s_x & s_x & -M_s & M_s & 0 \end{pmatrix} \Sigma \begin{pmatrix} l_x & -s_x \\ l_x & s_x \\ M_l & -M_s \\ M_l & M_s \\ 2 & 0 \end{pmatrix} \tag{91}$$

$$-\begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix} k(I - \tilde{\boldsymbol{\zeta}} \begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix}) = \begin{pmatrix} \psi & 0 \\ 0 & \psi \end{pmatrix} + \gamma \begin{pmatrix} \sigma_\beta^2 (l_x - M_l)^2 + 2\sigma_V^2 & 0 \\ 0 & \sigma_\beta^2 (s_x - M_s)^2 \end{pmatrix} \tag{92}$$

where $\tilde{\boldsymbol{\zeta}}$ is also diagonal and is equal to

$$\tilde{\boldsymbol{\zeta}} = \begin{pmatrix} \frac{1}{M_l} & 0 \\ 0 & \frac{1}{M_s} \end{pmatrix} \tag{93}$$

Hence l_x and s_x are decoupled and are defined by the following equations explicitly,

$$k \frac{l_x}{M_l}(l_x - M_l) = \psi + \gamma \left(\sigma_\beta^2 (l_x - M_l)^2 + 2\sigma_V^2 \right) \tag{94}$$

$$k \frac{s_x}{M_s}(s_x - M_s) = \psi + \gamma \sigma_\beta^2 (s_x - M_s)^2 \tag{95}$$

define the risk adjusted factors as

$$\phi_l = 1 - \frac{\gamma \sigma_\beta^2 M_l}{k} \qquad \phi_s = 1 - \frac{\gamma \sigma_\beta^2 M_s}{k} \tag{96}$$

The two equations above can be written as

$$\phi_l (M_l - l_x)^2 - M_l (M_l - l_x) - (\psi + 2\gamma \sigma_V^2) \frac{M_l}{k} = 0 \tag{97}$$

$$\phi_s (M_s - s_x)^2 - M_s (M_s - s_x) - \psi \frac{M_s}{k} = 0 \tag{98}$$

Hence

$$M_l - l_x = \frac{M_l(1 + \sqrt{1 + \frac{4\phi_l}{kM_l}(\psi + 2\gamma\sigma_V^2)})}{2\phi_l} \tag{99}$$

$$M_s - s_x = \frac{M_s(1 + \sqrt{1 + \frac{4\phi_s}{kM_s}\psi})}{2\phi_s} \tag{100}$$

We define $B_l \equiv M_l - l_x$ and $B_s \equiv M_s - s_x$, for ease of exposition in the calibration section.

To see how ψ affects the volatilities, define the instantaneous level and spread variances

$$V_s \equiv \frac{\mathrm{Var}(dP_{s,t})}{dt} = \frac{1}{2}\sigma_\beta^2(M_s - s_x)^2, \tag{101}$$

$$V_l \equiv \frac{\mathrm{Var}(dP_{l,t})}{dt} = \frac{1}{2}\left(\sigma_\beta^2(M_l - l_x)^2 + 2\sigma_V^2\right). \tag{102}$$

the ψ-derivatives are

$$\frac{\partial V_s}{\partial \psi} = \sigma_\beta^2(M_s - s_x)\frac{\partial(M_s - s_x)}{\partial \psi} \tag{103}$$

$$\frac{\partial V_l}{\partial \psi} = \sigma_\beta^2(M_l - l_x)\frac{\partial(M_l - l_x)}{\partial \psi} \tag{104}$$

From (99) and (100), it is straighforward to see that

$$\frac{\partial(M_s - s_x)}{\partial \psi} > 0 \qquad \frac{\partial(M_l - l_x)}{\partial \psi} > 0 \tag{105}$$

Higher capacity cost ψ therefore raises both level and spread volatility.

Similarly

$$\frac{\partial V_s}{\partial \gamma} = \sigma_\beta^2(M_s - s_x)\frac{\partial(M_s - s_x)}{\partial \gamma} \tag{106}$$

$$\frac{\partial V_l}{\partial \gamma} = \sigma_\beta^2(M_l - l_x)\frac{\partial(M_l - l_x)}{\partial \gamma} \tag{107}$$

Plug in $\frac{\partial(M_s-s_x)}{\partial\gamma}$ and $\frac{\partial(M_l-l_x)}{\partial\gamma}$ we have

$$\frac{\partial V_s}{\partial\gamma} = \frac{\sigma_\beta^4(M_s-s_x)^3}{k\sqrt{1+\frac{4\phi_s}{kM_s}\psi}} > 0 \tag{108}$$

$$\frac{\partial V_l}{\partial\gamma} = \frac{\sigma_\beta^2(M_l-l_x)(\sigma_\beta^2(M_l-l_x)^2+2\sigma_V^2)}{k\sqrt{1+\frac{4\phi_l}{kM_l}(\psi+2\gamma\sigma_V^2)}} > 0 \tag{109}$$

Hence higher risk aversion γ also raises both level and spread volatility.

A.5 Proof of Proposition 4

The volatility ratio of spread to level is given by

$$R \equiv \frac{V_s}{V_l} = \frac{\sigma_\beta^2(M_s-s_x)^2}{\sigma_\beta^2(M_l-l_x)^2+2\sigma_V^2} \tag{110}$$

Define $\Delta(a,\tilde{\psi})$ as the positive root to $\phi(a)\Delta^2 - a\Delta - \tilde{\psi}a/k = 0$, so that $\Delta_l \equiv M_l - l_x = \Delta(M_l, \psi+2\gamma\sigma_V^2)$ and $\Delta_s \equiv M_s - s_x = \Delta(M_s,\psi)$. Implicit differentiation gives

$$\frac{\partial\Delta}{\partial\tilde{\psi}} = \frac{1}{kD(a,\tilde{\psi})} > 0, \qquad \frac{\partial\Delta}{\partial a} = -\frac{\phi'(a)\Delta^2 - \Delta - \tilde{\psi}/k}{2\phi(a)\Delta - a} > 0, \tag{111}$$

where $D(a,\tilde{\psi}) \equiv \sqrt{1+\frac{4\phi(a)}{ka}\tilde{\psi}}$, $\phi(a) \equiv 1 - \frac{\gamma\sigma_\beta^2 a}{k}$, and $\phi'(a) = -\frac{\gamma\sigma_\beta^2}{k} < 0$. In the symmetric case with substitutable assets ($\boldsymbol{\zeta}_{12} < 0$), $M_l > M_s$ and the effective cost in the level equation, $\psi + 2\gamma\sigma_V^2$, strictly exceeds ψ. Monotonicity in both arguments therefore implies $\Delta_l > \Delta_s$. Combining with (101)–(102),

$$R = \frac{\sigma_\beta^2\Delta_s^2}{\sigma_\beta^2\Delta_l^2+2\sigma_V^2} < \frac{\sigma_\beta^2\Delta_l^2}{\sigma_\beta^2\Delta_l^2+2\sigma_V^2} < 1, \tag{112}$$

with strict inequality whenever $\sigma_V^2 > 0$. Hence $R < 1$.

Define $Q \equiv \frac{M_s-s_x}{M_l-l_x}$, we can write

$$R = \frac{Q^2}{1+\frac{2\sigma_V^2}{\sigma_\beta^2(M_l-l_x)^2}} \tag{113}$$

Taking log and derivative with respect to ψ we have

$$\frac{\partial \log R}{\partial \psi} = 2\frac{\partial \log Q}{\partial \psi} + \frac{4\sigma_V^2}{\sigma_\beta^2(M_l - l_x)^2 + 2\sigma_V^2} \underbrace{\frac{\partial \log(M_l - l_x)}{\partial \psi}}_{>0} \tag{114}$$

Furthermore,

$$\frac{\partial \log Q}{\partial \psi} = \frac{1}{M_s - s_x}\frac{\partial (M_s - s_x)}{\partial \psi} - \frac{1}{M_l - l_x}\frac{\partial (M_l - l_x)}{\partial \psi} \tag{115}$$

where

$$\frac{\partial (M_s - s_x)}{\partial \psi} = \frac{1}{k\sqrt{1 + \frac{4\phi_s}{kM_s}\psi}} \tag{116}$$

$$\frac{\partial (M_l - l_x)}{\partial \psi} = \frac{1}{k\sqrt{1 + \frac{4\phi_l}{kM_l}(\psi + 2\gamma\sigma_V^2)}} \tag{117}$$

$$\frac{1}{M_s - s_x}\frac{\partial (M_s - s_x)}{\partial \psi} = \frac{1}{k\sqrt{1 + \frac{4\phi_s}{kM_s}\psi}} \frac{2\phi_s}{M_s(1 + \sqrt{1 + \frac{4\phi_s}{kM_s}\psi})} \tag{118}$$

$$\frac{1}{M_l - l_x}\frac{\partial (M_l - l_x)}{\partial \psi} = \frac{1}{k\sqrt{1 + \frac{4\phi_l}{kM_l}(\psi + 2\gamma\sigma_V^2)}} \frac{2\phi_l}{M_l(1 + \sqrt{1 + \frac{4\phi_l}{kM_l}(\psi + 2\gamma\sigma_V^2)})} \tag{119}$$

Define

$$g(a, \tilde{\psi}) \equiv \frac{\phi(a)}{aD(a, \tilde{\psi})(1 + D(a, \tilde{\psi}))}, \qquad D(a, \tilde{\psi}) \equiv \sqrt{1 + \frac{4\phi(a)}{ka}\tilde{\psi}} \quad \phi(a) \equiv 1 - \frac{\gamma\sigma_\beta^2 a}{k} \tag{120}$$

Given $D(a, \tilde{\psi})$ is increasing in $\tilde{\psi}$, we have $g(a, \tilde{\psi})$ is decreasing in $\tilde{\psi}$. Furthermore, $g(a, \tilde{\psi})$ is decreasing in a. Since we have $M_l > M_s$, we have $g(M_l, \psi + 2\gamma\sigma_V^2) < g(M_s, \psi)$, which implies $\frac{1}{M_s - s_x}\frac{\partial (M_s - s_x)}{\partial \psi} > \frac{1}{M_l - l_x}\frac{\partial (M_l - l_x)}{\partial \psi}$. Hence $\frac{\partial \log Q}{\partial \psi} > 0$.

Similarly, taking derivative with respect to γ we have

$$\frac{\partial \log R}{\partial \gamma} = \frac{\partial \log V_s}{\partial \gamma} - \frac{\partial \log V_l}{\partial \gamma} \tag{121}$$

where

$$\frac{\partial \log V_s}{\partial \gamma} = \frac{1}{V_s}\frac{\partial V_s}{\partial \gamma} = \frac{2\sigma_\beta^2(M_s - s_x)}{k\sqrt{1+\frac{4\phi_s}{kM_s}\psi}} \tag{122}$$

$$\frac{\partial \log V_l}{\partial \gamma} = \frac{1}{V_l}\frac{\partial V_l}{\partial \gamma} = \frac{2\sigma_\beta^2(M_l - l_x)}{k\sqrt{1+\frac{4\phi_l}{kM_l}(\psi + 2\gamma\sigma_V^2)}} \tag{123}$$

Hence

$$\frac{\partial \log R}{\partial \gamma} = \frac{2\sigma_\beta^2}{k}\left(\frac{M_s - s_x}{\sqrt{1+\frac{4\phi_s}{kM_s}\psi}} - \frac{M_l - l_x}{\sqrt{1+\frac{4\phi_l}{kM_l}(\psi + 2\gamma\sigma_V^2)}}\right) \tag{124}$$

$$= \frac{2\sigma_\beta^2}{k}\left(\frac{M_s}{2\phi_s}\left(1 + \frac{1}{\sqrt{1+\frac{4\phi_s}{kM_s}\psi}}\right) - \frac{M_l}{2\phi_l}\left(1 + \frac{1}{\sqrt{1+\frac{4\phi_l}{kM_l}(\psi + 2\gamma\sigma_V^2)}}\right)\right) \tag{125}$$

we have

$$\frac{M_s}{2\phi_s}\left(1 + \frac{1}{\sqrt{1+\frac{4\phi_s}{kM_s}\psi}}\right) < \frac{M_s}{\phi_s} \tag{126}$$

and

$$\frac{M_l}{2\phi_l}\left(1 + \frac{1}{\sqrt{1+\frac{4\phi_l}{kM_l}(\psi + 2\gamma\sigma_V^2)}}\right) > \frac{M_l}{2\phi_l} \tag{127}$$

Under the condition that $\frac{2M_s}{\phi_s} < \frac{M_l}{\phi_l}$, we have $\frac{\partial \log R}{\partial \gamma} < 0$.

A.6 Proof of Corollary 2

By symmetry, $\mathrm{Var}(dP_{C,t}) = \mathrm{Var}(dP_{S,t}) \equiv \sigma_P^2$. The level and spread variances are

$$V_l = \frac{1}{4}\mathrm{Var}(dP_{C,t} + dP_{S,t}) = \frac{\sigma_P^2(1+\rho)}{2}, \quad V_s = \frac{1}{4}\mathrm{Var}(dP_{S,t} - dP_{C,t}) = \frac{\sigma_P^2(1-\rho)}{2} \tag{128}$$

where $\rho = \mathrm{Corr}(dP_{C,t}, dP_{S,t})$. Taking the ratio gives

$$R \equiv \frac{V_s}{V_l} = \frac{1-\rho}{1+\rho} \quad \Longleftrightarrow \quad \rho = \frac{1-R}{1+R} \tag{129}$$

Since ρ is strictly decreasing in R,

$$\frac{\partial \rho}{\partial \theta} = -\frac{2}{(1+R)^2}\frac{\partial R}{\partial \theta} \tag{130}$$

for any parameter θ. From Proposition 4, $\partial R/\partial \psi > 0$ implies $\partial \rho/\partial \psi < 0$, and $\partial R/\partial \gamma < 0$ (under the sufficient condition) implies $\partial \rho/\partial \gamma > 0$.

B Estimation Details

This appendix provides additional details on the estimation procedure described in Section 3.

B.1 Exponential Decay Approach

The polynomial decay approximation in equation (42) is a flexible but agnostic functional form. The model, however, implies a specific structure: because prices are affine in Ornstein-Uhlenbeck state variables, the impulse response function has an exponential decay form,

$$\theta_i(\tau) = \tilde{d}_{0,i} + \tilde{d}_{1,i}\, e^{-\nu_1 \tau} + \tilde{d}_{2,i}\, e^{-\nu_2 \tau}, \quad i = l, s, \tag{131}$$

where $\nu_1, \nu_2 > 0$ are decay rates (shared across level and spread) and $\tilde{d}_{0,i}, \tilde{d}_{1,i}, \tilde{d}_{2,i}$ are market-specific loading coefficients. The instantaneous decay rate under this specification is

$$\hat{m}_i = -\left(\nu_1\, \tilde{d}_{1,i} + \nu_2\, \tilde{d}_{2,i}\right). \tag{132}$$

We estimate the parameters in (131) using nonlinear least squares (NLS) applied to the local projection specification (41):

$$\Delta_\tau y_{i,t} = a_i + \left(\tilde{d}_{0,i} + \tilde{d}_{1,i}\, e^{-\nu_1 \tau} + \tilde{d}_{2,i}\, e^{-\nu_2 \tau}\right) \beta_{i,t} + c_{i,\tau}\, \Delta_\tau y_{i,t-\tau-1} + \varepsilon_{i,t+\tau}. \tag{133}$$

Two-step implementation. Since the nonlinearity is purely in the parameter space (the regressors are functions of parameters and the horizon τ, not of endogenous variables), we can reduce the dimensionality of the NLS problem using the Frisch-Waugh-Lovell (FWL) theorem. In the first step, we project out the control variables from both the dependent

variable and the shock:

$$\begin{aligned}\Delta_\tau y_{i,t} &= a_i^y + c_{i,\tau}^y \, \Delta_\tau y_{i,t-\tau-1} + \varepsilon_{i,t+\tau}^y, \\ \beta_{i,t} &= a_i^\beta + c_{i,\tau}^\beta \, \Delta_\tau y_{i,t-\tau-1} + \varepsilon_{i,t+\tau}^\beta,\end{aligned} \tag{134}$$

and collect the residuals $\Delta_\tau \check{y}_{i,t}$ and $\check{\beta}_{i,t}$. In the second step, we estimate the exponential model on the residuals via NLS:

$$\Delta_\tau \check{y}_{i,t} = \left(\tilde{d}_{0,i} + \tilde{d}_{1,i}\, e^{-\nu_1 \tau} + \tilde{d}_{2,i}\, e^{-\nu_2 \tau}\right) \check{\beta}_{i,t} + \varepsilon_{i,t+\tau}. \tag{135}$$

This substantially reduces the number of parameters in the nonlinear optimization, improving convergence.

B.2 Support Identification Algorithm

Our theory identifies intermediary costs from the instantaneous decay rate of the price impact. To estimate this decay rate, we need to identify the point at which the impulse response begins to decay—its turning point. This section describes the data-driven procedure for locating this turning point and setting the start of the estimation window $\underline{\tau}$ (referenced in Section 3.5). The end of the window is set to $\overline{\tau} = \underline{\tau} + w$, where w is a fixed window width (20 trading days in our baseline).

Step 0: Significance check. Test whether the impulse response is significantly different from zero at any horizon within the first 10 days, using the local projection estimates from (41) at the 5% significance level. If neither the level nor the spread response is significant, set the default start point $\underline{\tau} = 0$ and skip the remaining steps.

Step 1: High-order polynomial fit. Estimate a high-order polynomial (order 8 in our baseline) over the full horizon $[0, H]$ using the polynomial decay specification (42). Denote the fitted polynomial $\hat{f}(\tau)$.

Step 2: Critical point detection. Compute the numerical first derivative $\hat{f}'(\tau)$ and locate all zero-crossings (sign changes). For each zero-crossing that falls between two discrete horizons τ_1^* and τ_2^*, select the point $\tau_{\text{crit}}^* = \arg\min_{\tau_1^*, \tau_2^*} |\hat{f}'(\tau^*)|$.

Step 3: Classification. Compute the second derivative $\hat{f}''(\tau)$ at each critical point. Classify points where $\hat{f}'' < 0$ as peaks and points where $\hat{f}'' > 0$ as troughs.

Step 4: Start point selection. The goal of this step is to identify the turning point of the impulse response—the horizon at which the response peaks and begins to decay. We select the start of the estimation window as the peak of the fitted polynomial among $\tau = 0$ and all peak candidates identified in Step 3:

$$\underline{\tau} = \arg\max \left\{ |\hat{f}(0)|, |\hat{f}(\tau^*_{\text{peak},1})|, \ldots, |\hat{f}(\tau^*_{\text{peak},K})| \right\}.$$

The end of the estimation window is then $\overline{\tau} = \underline{\tau} + w$. By anchoring the window at this turning point, the decay estimation captures the transition from the peak response toward its long-run level, avoiding the pre-peak buildup phase where the response is still increasing.

Figure 10 illustrates this procedure for the baseline specification. The local projection estimates (blue circles) show the cumulative yield response rising over the first week before beginning to decline. The order-8 polynomial (grey dashed line), fitted over the full horizon, smooths through this noisy path and captures the turning point at Day 7—the peak of the fitted polynomial. The algorithm selects this as the start of the estimation window $\underline{\tau} = 7$, and the window extends $w = 20$ trading days to $\overline{\tau} = 27$ (green dashed lines). The linear decay function (dash-dotted line) is then fitted within this window, estimating the rate at which the price impact dissipates from its peak.

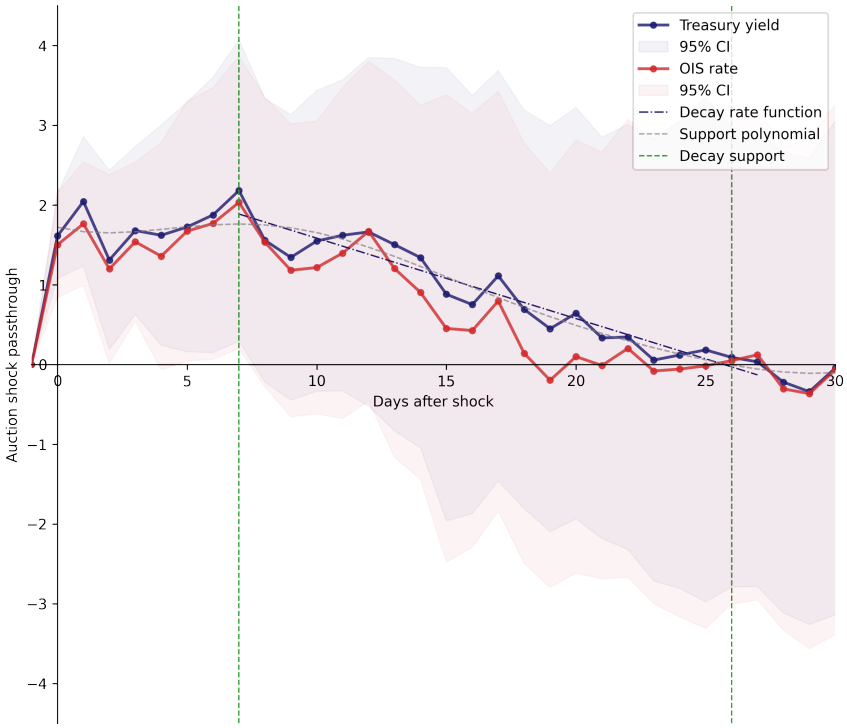


Figure 10: Support identification for the baseline specification

The figure shows the cumulative impulse response of Treasury yields (blue) and OIS rates (red) to a 5-year Treasury auction demand shock, with 95% confidence intervals. The grey dashed line is the order-8 polynomial fitted over the full horizon $[0, 30]$ used to detect the turning point. The green dashed lines mark the estimated decay window $[\underline{\tau}, \overline{\tau}] = [7, 27]$. The dash-dotted line shows the linear decay function fitted within this window. Full sample, 2008–2022.

Robustness to polynomial order. The turning point detection relies on the high-order polynomial fit in Step 1. Table 6 shows that the results are invariant to the polynomial order used for this step. For orders 6 through 10, the detected start point is Day 7 in every case, and the estimated decay rates and risk contributions are identical across all specifications. The 95% confidence interval for the yield's risk contribution excludes zero in all cases, confirming the dominance of risk-based costs. In practice, the polynomial order should be high enough to capture turning points in the impulse response—too low an order may smooth over the peak and misidentify the start of the decay phase—but not so high as to overfit noise in the local projection estimates. We use order 8 as our baseline.

Table 6: Robustness to polynomial order for start point detection

	Decay estimation	
	Start day	Window
Order = 6	7	20 days
Order = 7	7	20 days
Order = 8 (Baseline)	7	20 days
Order = 9	7	20 days
Order = 10	7	20 days

The table reports the detected start day and estimation window for the baseline specification (5-year Treasury note auction demand shock, full sample 2008–2022) under different polynomial orders used for turning point detection in Step 1 of the support identification algorithm (Section B.2). All other parameters are held at their baseline values.

B.3 Constrained Estimation and Bootstrap Inference

The requirement that risk contributions lie in $[0, 1]$ imposes bounds on the ratio of decay rates:

$$\frac{m_s}{m_l} \in \left[\frac{\sigma_{1,s}}{\sigma_1^2},\, 1\right]. \tag{136}$$

The lower bound equals $1 - \beta_{1,2}$, where $\beta_{1,2}$ is the coefficient from regressing Treasury yield changes on OIS rate changes. Under exact no-arbitrage ($\beta_{1,2} = 1$), the lower bound is zero—the spread does not respond at all. The upper bound of one requires that the spread decays no faster than the level.

When computing the risk contribution estimates $\hat{C}_{r,l}$ and $\hat{C}_{r,s}$ from equations (39)–(40), we impose the theoretical bounds (136) to ensure that the estimated contributions are economically meaningful ($\hat{C}_{r,l}, \hat{C}_{r,s} \in [0, 1]$). This requires that the ratio m_s/m_l lie in $[\sigma_{1,s}/\sigma_1^2, 1]$.

Constrained optimization. We first project out control variables (intercepts and lagged dependent variables) from both the dependent variable and shock regressors using OLS, following the FWL theorem. Working with the projected residuals, we solve the constrained

least squares problem

$$\min_{\mathbf{b}} \frac{1}{2}\mathbf{b}^\top(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})\mathbf{b} - (\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}})^\top\mathbf{b} \quad \text{subject to} \quad \frac{m_s}{m_l} \in \left[\frac{\sigma_{1,s}}{\sigma_1^2}, 1\right], \tag{137}$$

where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ are the FWL-projected shock regressors and dependent variable, $\mathbf{b}$ is the vector of polynomial coefficients, and $m_l = b_{l,1}$, $m_s = b_{s,1}$ are the linear coefficients. The constraints are implemented as inequality constraints on the elements of $\mathbf{b}$ and solved via sequential quadratic programming (SLSQP).

Bootstrap inference. We use a date-cluster bootstrap for inference: in each of 1,000 replications, we resample trading dates with replacement (preserving the within-date correlation structure), re-estimate the constrained problem on the bootstrap sample, and record the resulting estimates. The standard errors for $\hat{m}_l$ and $\hat{m}_s$ are computed as the standard deviation of the bootstrap distribution. Standard errors for the risk contributions $\hat{C}_{r,l}$ and $\hat{C}_{r,s}$ are obtained via the delta method applied to the bootstrap covariance matrix of the slope estimates.

R^2 loss diagnostic. To assess the cost of imposing the theoretical bounds, we report the R^2 loss: the difference in R^2 between the unconstrained and constrained fits. A small R^2 loss indicates that the constraint is nearly satisfied by the data and the constrained estimates are close to the unrestricted optimum.

Table 7 reports this diagnostic for the three baseline specifications alongside the constrained risk contribution estimates. For the full sample, the constraint binds with an R^2 loss of 3.15% of the unconstrained R^2. For the GFC, the constraint also binds with an R^2 loss of 1.42%. For COVID, the constraint binds but the R^2 loss is negligible (0.01%). The small R^2 losses reflect the fact that the theoretical bounds depend on the ratio of the spread and level decay slopes. When the spread slope is near zero—as in the full sample and GFC, where the spread shows no statistically significant response—small estimation noise can push the unconstrained ratio slightly outside the admissible region, but the correction required to restore feasibility is minimal.

Table 7: R^2 loss from imposing theoretical bounds

	Risk contribution (%)		Constrained estimation	
	Yield	Spread	Constraint binds	R^2 loss (%)
Full sample	100.0***	100.0***	Yes	3.15
	(2.3)	(20.3)		
GFC	100.0***	100.0**	Yes	1.42
	(4.9)	(40.5)		
COVID	0.0	0.0	Yes	0.01
	(100.4)	(20.3)		

The table reports risk contribution estimates and the R^2 loss from imposing the theoretical bounds on the ratio of decay rates (equation (136)). Standard errors use date-cluster bootstrap (1,000 replications); risk contribution standard errors use the delta method. R^2 loss is expressed as a percentage of the unconstrained R^2, computed on the Frisch–Waugh–Lovell residuals of the joint polynomial decay regression. All specifications use a linear decay function with a 20-day estimation window. Significance: *** p<0.01, ** p<0.05, * p<0.1.

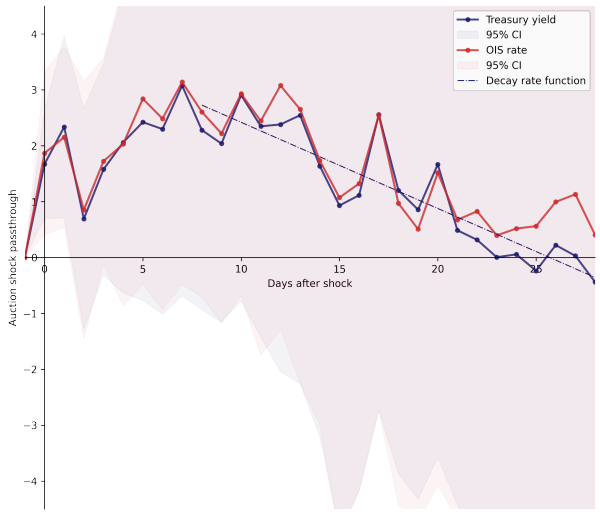
C Crisis Episode Impulse Responses: Levels

Figure 11 shows the impulse responses of Treasury yields and the OIS rate to 5-year Treasury note auction demand shocks during the two crisis episodes.

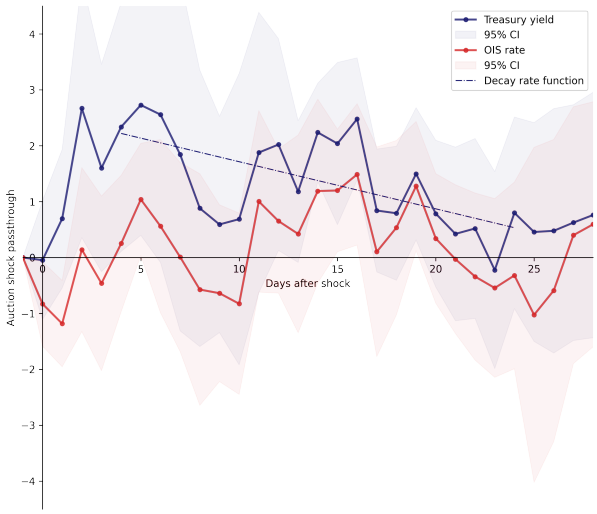
Tables 8 and 9 report robustness of the crisis episode estimation results across alternative specifications.

Figure 11: Level impulse response to Treasury auction demand shocks: Crisis episodes

(a) GFC (Jan 2008 to Jan 2010)



(b) COVID (Feb 2020 to Feb 2021)



The figures plot impulse response functions of 5-year Treasury yields (blue) and the OIS rate (red) to 5-year Treasury note auction demand shocks for two crisis episodes, with 95% confidence bands. The dashed line shows the fitted decay function. Panel (a) shows the GFC episode (January 2008 to January 2010). Panel (b) shows the COVID episode (February 2020 to July 2020).

Table 8: Robustness: GFC episode (Jan 2008 to Jan 2010)

	Risk contribution (95% CI)		Instantaneous decay (pp)		Decay estimation	
	Yield	Spread	Yield	Spread	function	Window
Baseline	**90% – 100%**	**21% – 100%**	-0.151 (0.109)	-0.018 (0.014)	Linear	20 days
Window (15d)	**72% – 100%**	0% – 100%	-0.152 (0.127)	-0.018 (0.024)	Linear	15 days
Window (10d)	**49% – 100%**	0% – 100%	-0.111 (0.167)	-0.013 (0.035)	Linear	10 days
Quadratic	0% – 100%	0% – 100%	-0.009 (0.205)	-0.001 (0.028)	Quadratic	20 days

The table reports instantaneous decay rates and risk contribution estimates for Treasury auction demand shocks during the GFC episode (January 2008 to January 2010) across different specifications. The demand shock is constructed from 5-year Treasury note auctions. Instantaneous decay rates are in percentage points per day. Risk contribution decomposes the yield response into level (risk-based) and spread (gross position) costs. Standard errors use date-cluster bootstrap (1,000 replications) on constrained estimates. Risk contribution confidence intervals use the delta method, bounded to $[0, 100]$. Bold confidence intervals indicate statistically informative decomposition of risk-based versus gross position

Table 9: Robustness: COVID episode (Feb 2020 to Feb 2021)

	Risk contribution (95% CI)		Instantaneous decay (pp)		Decay estimation	
	Yield	Spread	Yield	Spread	function	Window
Baseline	0% – 100%	**0% – 40%**	-0.087 (0.133)	-0.087 (0.075)	Linear	20 days
Window (15d)	0% – 100%	**0% – 41%**	-0.075 (0.115)	-0.075 (0.069)	Linear	15 days
Window (10d)	0% – 100%	**0% – 77%**	-0.097 (0.218)	-0.097 (0.095)	Linear	10 days
Quadratic	0% – 100%	0% – 100%	-0.088 (0.280)	-0.088 (0.143)	Quadratic	20 days

The table reports instantaneous decay rates and risk contribution estimates for Treasury auction demand shocks during the COVID episode (February 2020 to February 2021) across different specifications. The demand shock is constructed from 5-year Treasury note auctions. Instantaneous decay rates are in percentage points per day. Risk contribution decomposes the yield response into level (risk-based) and spread (gross position) costs. Standard errors use date-cluster bootstrap (1,000 replications) on constrained estimates. Risk contribution confidence intervals use the delta method, bounded to $[0, 100]$. Bold confidence intervals indicate statistically informative decomposition of risk-based versus gross position costs. Significance: *** p<0.01, ** p<0.05, * p<0.1.