Which Intermediary Costs Matter for Asset Prices?*

Manav Chaudhary† Julie Zhiyu Fu‡ Jian Li§

February 27, 2026

Abstract

Arbitrage spreads, such as the Treasury–OIS spread, are a standard diagnostic for intermediary capacity. However, because spreads are relative objects, they may miss the intermediation costs that govern the level of asset prices, such as Treasury yields. We develop a model for the joint dynamics of price levels and spreads where intermediaries face two costs: risk-based costs for bearing portfolio risk and gross position costs tied to position size. The latter primarily affect spreads, while the former primarily drive price levels. We show that differences in the initial rate at which prices and spreads revert following a demand shock separately identify each type of cost. We take this framework to the U.S. Treasury cash and OIS markets. Using high-frequency shocks from U.S. Treasury auctions, we find that risk-based costs dominate on average: yields comove with OIS rates, leaving spreads unchanged. The exception is the COVID crisis, where gross position costs dominate, consistent with the introduction of the supplementary leverage ratio (SLR). Consequently, while spreads are an effective diagnostic during acute stress episodes in the post-SLR regime, they generally offer a limited view of the intermediation costs that drive price-level dynamics in typical market environments.

*We thank Daniel Barth, Dimitri Vayanos and seminar participants at the Demand in Asset Markets workshop, London Junior Conference 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

Spreads between nearly identical assets, such as the Treasury-OIS spread, are widely used as evidence that intermediary frictions are distorting capital markets. Policymakers often interpret elevated spreads as signals of market dysfunction (Board of Governors of the Federal Reserve System, 2020). When a near-riskless arbitrage relationship "breaks," it is taken as a sign that intermediation capacity has become scarce and that liquidity provision is impaired. A large literature on intermediary asset pricing formalizes this view, emphasizing that intermediary balance sheet capacity is a first-order determinant of equilibrium prices (He and Krishnamurthy, 2013; Brunnermeier and Sannikov, 2014). Some of the most influential empirical evidence for this view traces the effects of intermediary capacity precisely through the behavior of arbitrage spreads (Du, Tepper, and Verdelhan, 2018; Du, Hébert, and Huber, 2023; D. Duffie et al., 2023).[1]

At the same time, focusing on spreads may offer only a partial view of whether intermediary balance sheet capacity is affecting capital markets. Spreads are inherently relative objects: they capture price differences across assets that should be linked by no-arbitrage. As such, they are natural diagnostics for frictions that impede intermediaries from conducting relative arbitrage. But many of the welfare- and policy-relevant outcomes—government borrowing costs, discount rates, the transmission of monetary policy, and the pricing of safe assets—are governed by movements in the level of asset prices. These price levels, such as the Treasury yield, are not pinned down by riskless relative arbitrage. A broader set of frictions may affect them while leaving only a limited trace in spreads. To what extent, then, are spreads informative about distortions in the level of prices themselves? Answering this question requires a framework that models spreads and price levels jointly.

This paper develops such a framework. We build a preferred-habitat 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 arbitrage spreads, risk-based costs are the dominant force behind movements in price levels. This distinction underscores the importance of separately identifying the two types of costs.

We develop a new identification strategy leveraging tools from demand-system asset pric-

[1] Papers such as He, Kelly, and Manela (2017) and Haddad and Muir (2025) study the link between intermediary balance sheets and aggregate asset prices, but do not provide sharp identification.

ing. 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 the decay to the marginal cost of intermediation—it is, in effect, a sufficient statistic for intermediary costs. Because the two costs affect price levels and spreads differently, they induce distinct decay rates, 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 intermediary 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 can shift during crises. During the 2008 Global Financial Crisis, risk-based costs are the dominant driver of intermediary capacity; during the COVID crisis, gross position costs are. This shift is consistent with the post-2017 implementation of the supplementary leverage ratio, 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 intermediary 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 (J. D. Duffie, 2023). Taken together, these findings indicate that the relevant intermediary costs have evolved over time. As a result, the informativeness of spreads about price-level distortions has shifted as well.

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 and Section 4 presents the empirical results. Section 5 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

[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).

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 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

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

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).

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 $\mathbf{p} = (\boldsymbol{\lambda}_{\boldsymbol{x}}, \boldsymbol{\lambda}_{\boldsymbol{\beta}}, \boldsymbol{\lambda}_V)$ where $\boldsymbol{\lambda}_{\boldsymbol{x}}$ (2×2), $\boldsymbol{\lambda}_{\boldsymbol{\beta}}$ (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 (49) in Appendix A.

Define

$$\mathbf{\Lambda} \equiv k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{x}}) \tag{10}$$

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

$\boldsymbol{\lambda_x}$ 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*

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

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* (65) *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

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

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

$$\begin{pmatrix} m_l & 0 \\ 0 & m_s \end{pmatrix} \equiv \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \boldsymbol{\zeta}^{-1} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\top} \tag{27}$$

where m_l and m_s can be interpreted as the multiplier of average price level and 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. Consider the price impact dynamics in the cash market in response to a demand shock in the cash market. The long-run impact is determined by the institutional investors' demand elasticity, since eventually they step in and absorb the entire demand shock. However, in the short- and medium-run, the market relies on the intermediaries to absorb the demand shock, and they require return compensation to hold the extra inventory. As the institutional investors step in, the extra inventory held by intermediaries decline, hence the price impact starts at $-\lambda_{x,11} + \zeta_{11}^{-1}$ on impact and decays toward the long-run level ζ_{11}^{-1}. At each instant, the return the intermediaries get is the slope of the price impact curve, which is equal to the intermediary share times their marginal cost. On impact of the shock at $\tau = 0$, the intermediary share is 100%, and hence the slope of the price impact curve at $\tau = 0$ directly reveals the marginal cost of the intermediaries.

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.

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.

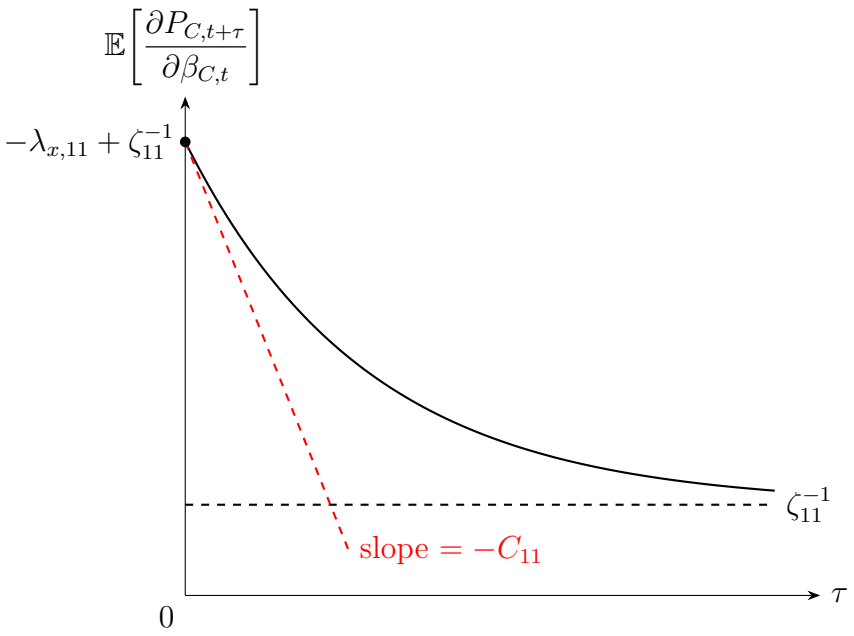


Figure 1: Price impact dynamics following a demand shock. The solid curve shows the expected price response $\mathbb{E}[\partial P_{C,t+\tau}/\partial \beta_{C,t}]$ as a function of the horizon τ. The price impact starts at $-\lambda_{x,11} + \zeta_{11}^{-1}$ on impact and decays toward the long-run level ζ_{11}^{-1} (dashed horizontal line). The slope at $\tau = 0$ equals $-C_{11}$, the negative of the intermediary's marginal cost.

3.2 Treasury Yields and OIS Rates

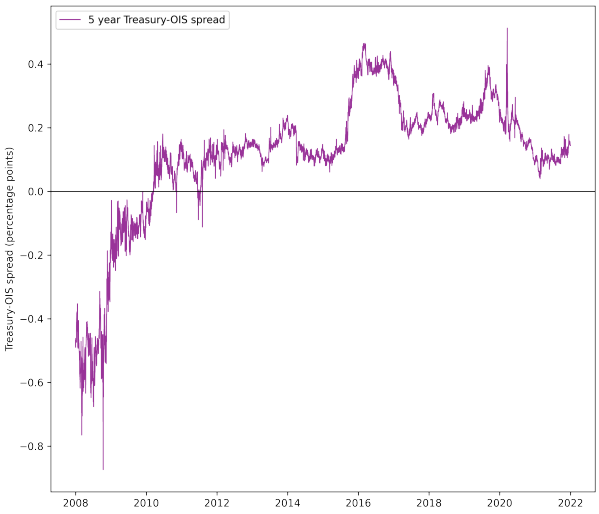
We measure the price level response using the 5-year constant-maturity Treasury yield, sourced from the U.S. Department of the Treasury's daily yield curve data. For the synthetic price level, we use the 5-year overnight index swap (OIS) rate from Bloomberg, where the underlying reference rate is the 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

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

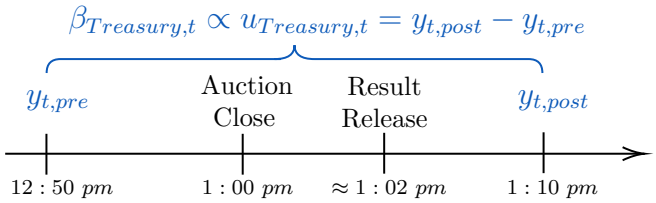
The intraday yield data are constructed from tick-level quotes sourced from GovPX, which provides real-time prices for on-the-run and when-issued Treasury securities. On non-auction days, the demand shock is set to zero.

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

[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.

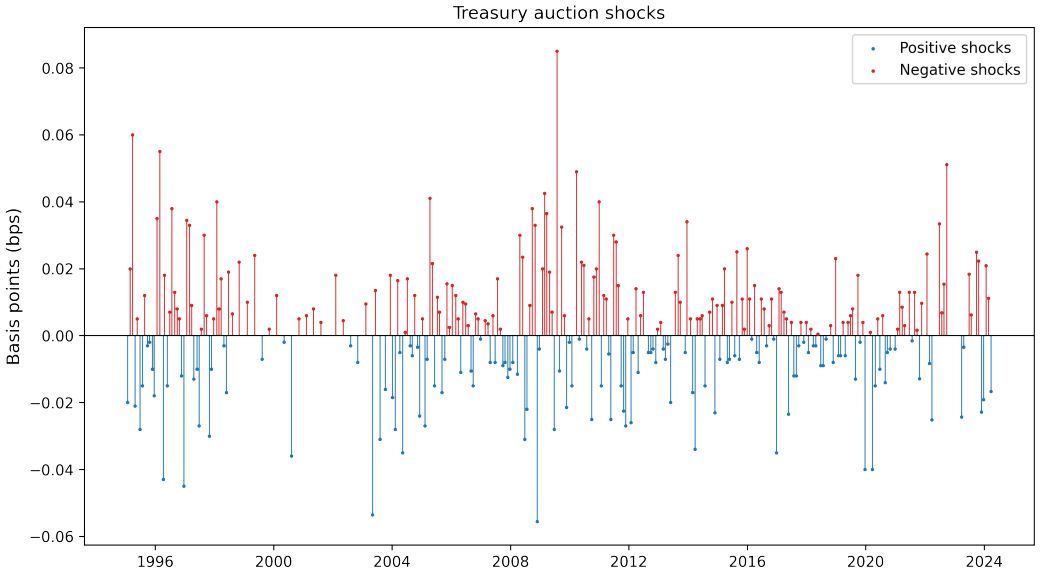
Figure 3: Timeline of Treasury auction shock construction



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

bond prices up and hence yields down, so that $u_{\text{Treasury},t} < 0$.

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 Concerns

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 are exogenous to macroeconomic fundamentals, and that the shock hits the Treasury market but not the OIS market.

Two concerns are particularly relevant for our identification. First, the auction demand shock could reflect the revelation of fundamental information. The high-frequency window of only a few minutes helps rule out contamination from macroeconomic news arriving simultaneously. A subtler concern is that auction participants may possess private information, so that their demand reveals fundamentals rather than reflecting an exogenous shift. Droste, Gorodnichenko, and Ray (2024) provide evidence against this channel: the surprise movements in 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 demand shock in the Treasury market may be correlated with a simultaneous demand shock in the OIS market, which would confound our identification of separate yield and spread responses. Our identification requires that the shock only hits the Treasury market directly; any subsequent spillover into OIS rates through arbitrage activity is part of the transmission mechanism we aim to capture. This assumption is plausible because the shock originates predominantly from foreign official sector participants, who operate directly in the cash Treasury market rather than in interest rate derivatives (see Droste, Gorodnichenko, and Ray, 2024). Foreign central banks and sovereign wealth funds typically build Treasury positions through outright purchases rather than via OIS or other derivative instruments.

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 Montiel Olea and Plagborg-Møller (2021).

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. However, during the COVID crisis, we find that no-arbitrage spreads respond strongly: all of the yield movement is explained by gross position costs rather than risk-based costs. Our findings suggest that post-GFC regulations imposing costs on gross position sizes become the main driver of intermediation capacity during periods of market stress.

4.1 Baseline Results

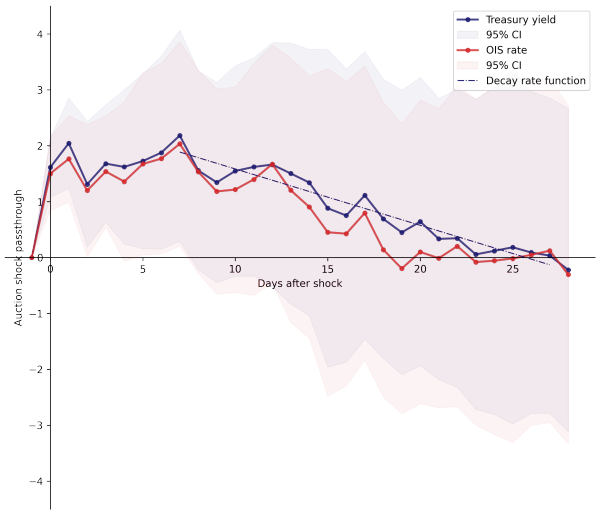
Figure 5 shows the impulse response functions to Treasury auction demand shocks for the full sample (2008–2022). Panel (a) shows that Treasury yields and the OIS rate move in lockstep: both initially increase and then slowly decay toward zero. This is consistent with slow-moving investors entering the 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.

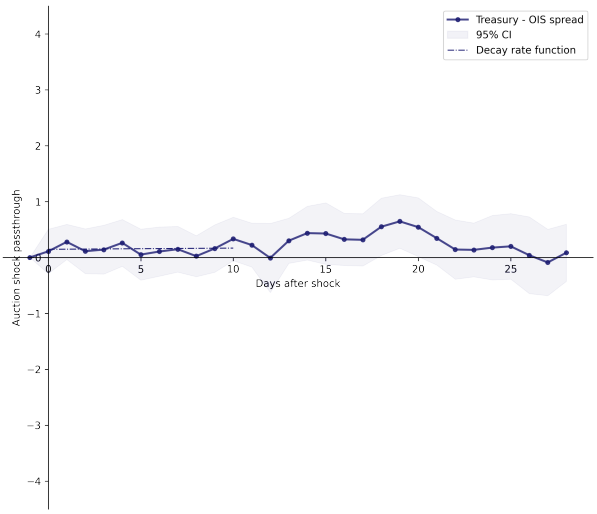
Table 3 presents the estimation results for the 5-year note auction demand shock across several specifications. The first two columns report 95% confidence intervals for the risk contribution decomposition of the yield response into level (risk-based) and spread (gross position) costs. The next two columns report instantaneous decay rates—the slope of the fitted decay function evaluated at $h = 0$—for the Treasury yield level and the Treasury-OIS

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.

spread, respectively.[11]

The baseline specification uses a linear decay function estimated over a 20-day window. The yield decay rate is -0.101 percentage points per day, significant at the 10% level, indicating a gradual but statistically detectable reversion of the yield effect. The spread decay rate, in contrast, is essentially zero (0.002, with a standard error of 0.025) and statistically indistinguishable from zero. This confirms that the Treasury-OIS spread does not respond to auction demand shocks, consistent with the visual evidence in Figure 5.

The risk contribution columns formalize this result. The 95% confidence interval for the yield's risk contribution is 79% to 100%, meaning that the entire yield response can be attributed to risk-based costs. The spread's risk contribution is imprecisely estimated,

[11]The estimation window for the decay function is automatically detected using a polynomial support identification approach described in Section 3.5. The shorter window specifications use the auto-detected start point but manually adjust the end point.

with wide confidence intervals reflecting the fact that the spread itself shows no meaningful response.

These findings are robust across alternative specifications. Shortening the estimation window to 15 or 10 days produces yield decay rates of -0.107 and -0.095 percentage points per day, respectively; although we lose power with smaller estimation windows, all point estimates are close to the baseline estimate. The 95% confidence interval for the yield's risk contribution excludes zero across all 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 yields a yield decay rate of -0.135 percentage points per day and the same risk contribution pattern. Finally, broadening the demand shock definition to include 3-year through 7-year Treasury auctions produces a yield decay rate of -0.072 percentage points per day with a risk contribution of 100% for the yield. The smaller magnitude is consistent with the transmission of shocks across the term structure of yields being imperfect, attenuating the 5-year yield response to the broad demand shock.

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

	Instantaneous decay (pp)		Risk contribution (%)		Decay estimation	
	Yield	Spread	Yield	Spread	function	Window
Baseline	-0.101*	0.002	100***	100	Linear	20 days
	(0.056)	(0.025)	(10.9)	(96.6)		
Window (15d)	-0.107*	0.013	100***	100***	Linear	15 days
	(0.062)	(0.019)	(3.9)	(34.3)		
Window (10d)	-0.095	0.002	100***	100	Linear	10 days
	(0.102)	(0.025)	(11.4)	(100.6)		
Window (5d)	-0.063	-0.009	97	77	Linear	5 days
	(0.146)	(0.046)	(89.0)	(495.2)		
Quadratic	-0.135	0.002	100***	100	Quadratic	20 days
	(0.200)	(0.025)	(8.8)	(77.6)		
Broad shock	-0.072	0.004	100**	100	Linear	4 days
	(0.099)	(0.009)	(44.2)	(391.4)		

Notes: Standard errors in parentheses. Significance: *** p<0.01, ** p<0.05, * p<0.1.

The table reports instantaneous decay rates and risk contribution estimates for Treasury auction demand shocks across different specifications. The sample spans $N = 3{,}621$ trading days from January 2008 to January 2022, with 156 auction dates. The baseline demand shock is constructed from 5-year Treasury note auctions. The broad shock definition aggregates 3-year, 5-year, and 7-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 in parentheses for decay rates use date-clustered OLS. Risk contributions report 95% confidence intervals bounded to $[0, 100]$, using date-cluster bootstrap standard errors. Bold confidence intervals indicate statistically informative decomposition of risk-based versus gross position costs.

4.2 The Story of Two Crises

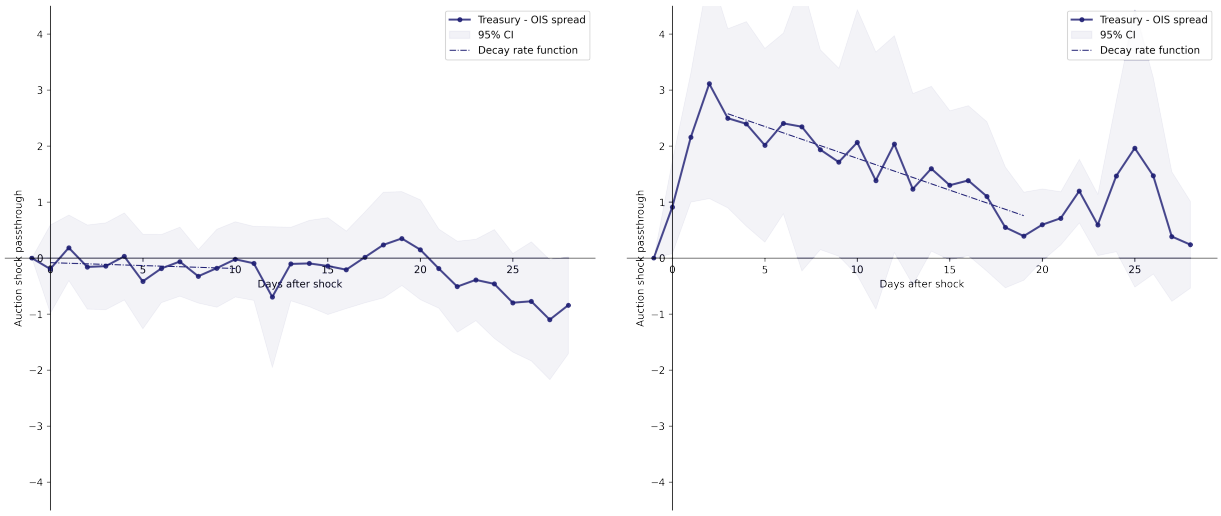
We now examine two crisis episodes—the Global Financial Crisis (GFC, 2008–2010) and the COVID-19 crisis (2020)—to test whether the dominant cost inhibiting intermediation has changed following the introduction of gross position regulatory constraints. In 2017, the Supplementary Leverage Ratio, enacted 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 starkly different picture for the COVID crisis: the spread jumps sharply on impact and decays gradually over the following weeks. This indicates that gross position costs—rather than risk-based costs—became the dominant driver of 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 became binding, causing the spread to respond.

Table 4 confirms this interpretation quantitatively. During the GFC, the yield decay rate is -0.157 percentage points per day, and the 95% confidence interval for the yield's risk contribution is 64% to 100%—indicating that risk-based costs account for the majority of the yield response. During the COVID crisis, the yield decay rate is -0.075 percentage points per day, but the yield's risk contribution CI spans 0% to 100%, indicating that the data cannot distinguish risk-based from gross position costs through the yield alone. The spread, however, is highly informative: the spread decay rate during COVID is -0.109 percentage points per day, significant at the 1% level, and the spread's risk contribution CI is bounded to 0% to 37%—indicating that gross position costs account for the majority of the spread response. This confirms the visible decay pattern in Figure 6b and the qualitative contrast with the GFC, where risk-based costs dominate.

Table 4: Estimation results: Crisis episodes

	Instantaneous decay (pp)		Risk contribution (%)		Decay estimation	
	Yield	Spread	Yield	Spread	function	Window
GFC	-0.157	-0.010	100***	100	Linear	16 days
	(0.118)	(0.048)	(18.4)	(152.4)		
COVID	-0.063	-0.114***	-0	-0	Linear	17 days
	(0.086)	(0.044)	(319.8)	(70.5)		

Notes: Standard errors in parentheses. Significance: *** p<0.01, ** p<0.05, * p<0.1.

The table reports instantaneous decay rates and risk contribution estimates for Treasury auction demand shocks during two crisis episodes. The GFC sample spans January 2008 to January 2010; the COVID sample spans February 2020 to February 2021. 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 in parentheses for decay rates use date-clustered OLS. Risk contributions report 95% confidence intervals bounded to $[0, 100]$, using date-cluster bootstrap standard errors. Bold confidence intervals indicate statistically informative decomposition of risk-based versus gross position costs. Impulse responses for Treasury yields and OIS rates are reported in Appendix C.

5 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 identification strategy exploits the intermediaries' first-order condition: the initial rate at which prices revert 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 on average, 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 binding constraint, consistent with the post-2017 implementation of the Supplementary Leverage Ratio imposing balance sheet costs on dealer positions.

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.

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.

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.

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.

— (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.

Montiel Olea, José Luis and Mikkel Plagborg-Møller (2021). "Local Projection Inference Is Simpler and More Robust than You Think". In: *Econometrica : journal of the Econometric Society* 89.4, pp. 1789–1823. DOI: 10.3982/ECTA18756.

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

$$\mathbf{x}_t + \mathbf{y}_t + \boldsymbol{\beta}_t = \bar{\mathbf{S}} \tag{43}$$

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

$$dy_{i,t} = k(Z_{i,t} - y_{i,t})dt \tag{44}$$

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

$$d\mathbf{y}_t = k(-\boldsymbol{\zeta}\mathbf{P}_t + V_t\boldsymbol{\zeta}\mathbf{1} + \boldsymbol{\theta} - \mathbf{y}_t)dt \tag{45}$$

substituting the market clearing condition we have

$$d\mathbf{x}_t = -k\Big(\mathbf{x}_t - \boldsymbol{\zeta}\mathbf{P}_t + V_t\boldsymbol{\zeta}\mathbf{1} + \boldsymbol{\theta} - \bar{\mathbf{S}} + \boldsymbol{\beta}_t\Big)dt - d\boldsymbol{\beta}_t. \tag{46}$$

Next, substitute (8) in (67), we get,

$$d\mathbf{x}_t = -k\Big(\mathbf{x}_t - \boldsymbol{\zeta}\mathbf{p}\mathbf{s}_t - \boldsymbol{\zeta}\bar{\mathbf{P}} + V_t\boldsymbol{\zeta}\mathbf{1} + \boldsymbol{\theta} - \bar{\mathbf{S}} + \boldsymbol{\beta}_t\Big)dt - d\boldsymbol{\beta}_t. \tag{47}$$

$$= -k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{x}})\mathbf{x}_t dt - k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}})\boldsymbol{\beta}_t dt + k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1})V_t dt + \text{constant}dt - d\boldsymbol{\beta}_t. \tag{48}$$

This implies the instantaneous innovation covariance matrix is given by

$$\Sigma \equiv \frac{\text{Cov}(d\mathbf{s}_t)}{dt} = \begin{pmatrix} \sigma_C^2 & \sigma_{CS} & -\sigma_C^2 & -\sigma_{CS} & -\sigma_{CV} \\ \sigma_{CS} & \sigma_S^2 & -\sigma_{CS} & -\sigma_S^2 & -\sigma_{SV} \\ -\sigma_C^2 & -\sigma_{CS} & \sigma_C^2 & \sigma_{CS} & \sigma_{CV} \\ -\sigma_{CS} & -\sigma_S^2 & \sigma_{CS} & \sigma_S^2 & \sigma_{SV} \\ -\sigma_{CV} & -\sigma_{SV} & \sigma_{CV} & \sigma_{SV} & \sigma_V^2 \end{pmatrix} \tag{49}$$

and the mean reversion matrix Γ is given by

$$\Gamma = \begin{pmatrix} \boldsymbol{\Lambda} & k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}}) - \boldsymbol{\eta}_\beta & -k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) \\ \mathbf{0}_{2\times 2} & \boldsymbol{\eta}_\beta & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 2} & \eta_v \end{pmatrix} \tag{50}$$

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

$$-\mathbf{p}\Gamma(\mathbf{s}_t - \bar{\mathbf{s}}) = \mathbf{C}\mathbf{x}_t \tag{51}$$

$$-\begin{pmatrix} \boldsymbol{\lambda}_{\boldsymbol{x}} & \boldsymbol{\lambda}_{\boldsymbol{\beta}} & \boldsymbol{\lambda}_V \end{pmatrix} \begin{pmatrix} \mathbf{\Lambda} & k(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}}) & -k\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) \\ \mathbf{0}_{2\times 2} & \boldsymbol{\eta}_\beta & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 2} & \eta_v \end{pmatrix} \begin{pmatrix} \mathbf{x}_t - \bar{\mathbf{x}} \\ \boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}} \\ V_t - \bar{V} \end{pmatrix} = \mathbf{C}\mathbf{x}_t \tag{52}$$

Expanding the equation we get

$$-\boldsymbol{\lambda}_{\boldsymbol{x}}\mathbf{\Lambda}(\mathbf{x}_t - \bar{\mathbf{x}}) - \boldsymbol{\lambda}_{\boldsymbol{\beta}}\boldsymbol{\eta}_\beta(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}) - \boldsymbol{\lambda}_V\eta_v(V_t - \bar{V}) - k\boldsymbol{\lambda}_{\boldsymbol{x}}(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}})(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}) \tag{53}$$

$$+ k\boldsymbol{\lambda}_{\boldsymbol{x}}\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1})(V_t - \bar{V}) = \mathbf{C}\mathbf{x}_t \tag{54}$$

Matching the coefficients we have

$$[\mathbf{x}_t]: \quad -\boldsymbol{\lambda}_{\boldsymbol{x}}\mathbf{\Lambda} = \mathbf{C} \tag{55}$$

$$[\boldsymbol{\beta}_t]: \quad -\boldsymbol{\lambda}_{\boldsymbol{\beta}}\boldsymbol{\eta}_\beta - k\boldsymbol{\lambda}_{\boldsymbol{x}}(I - \boldsymbol{\zeta}\boldsymbol{\lambda}_{\boldsymbol{\beta}}) = \mathbf{0} \tag{56}$$

$$[V_t]: \quad -\boldsymbol{\lambda}_V\eta_v + k\boldsymbol{\lambda}_{\boldsymbol{x}}\boldsymbol{\zeta}(\boldsymbol{\lambda}_V - \mathbf{1}) = \mathbf{0} \tag{57}$$

Consider the limiting case where $\eta_{\beta_i} \to 0$ and $\eta_v \to 0$. When $\boldsymbol{\lambda}_{\boldsymbol{x}}$ has full rank, from (56) and (57) we must have

$$\boldsymbol{\lambda}_{\boldsymbol{\beta}} = \boldsymbol{\zeta}^{-1} \tag{58}$$

$$\boldsymbol{\lambda}_V = \mathbf{1} \tag{59}$$

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 (55)

$$-\boldsymbol{\lambda}_{\boldsymbol{x}}\mathbf{\Lambda} = \begin{pmatrix} \psi_C & 0 \\ 0 & \psi_S \end{pmatrix} + \gamma \begin{pmatrix} \boldsymbol{\lambda}_{\boldsymbol{x}} & \boldsymbol{\zeta}^{-1} & \mathbf{1} \end{pmatrix} \Sigma \begin{pmatrix} \boldsymbol{\lambda}_{\boldsymbol{x}} \\ \boldsymbol{\zeta}^{-1} \\ \mathbf{1} \end{pmatrix} \tag{60}$$

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

$$\bar{\mathbf{P}} = \boldsymbol{\zeta}^{-1}\left(\boldsymbol{\theta} - \bar{\mathbf{S}}\right) \tag{61}$$

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 (55), we have

$$\boldsymbol{\lambda}_{\boldsymbol{x}} = -c\mathbf{1}\mathbf{1}^\top\boldsymbol{\Lambda}^{-1} \tag{62}$$

$$(1,-1)^\top\boldsymbol{\lambda}_{\boldsymbol{x}} = 0 \quad \Rightarrow \quad \lambda_{C,\mathbf{x}} = \lambda_{S,\mathbf{x}} = \lambda_{\mathbf{x}} \tag{63}$$

Plug this into (56), we have

$$\boldsymbol{\lambda}_{\boldsymbol{\beta}} = \frac{\lambda_{\mathbf{x}}}{\lambda_{\mathbf{x}}\boldsymbol{\zeta}(1,1)^\top} \tag{64}$$

From (57) we still have $\boldsymbol{\lambda}_V = \mathbf{1}$. Plug the coefficients back into (55) and multiply both sides by $\begin{pmatrix}1 & 0\end{pmatrix}$ we get

$$-\lambda_{\mathbf{x}}k(I - \boldsymbol{\zeta}\begin{pmatrix}1\\1\end{pmatrix}\lambda_{\mathbf{x}}) = \gamma\begin{pmatrix}\lambda_{\mathbf{x}} & \frac{\lambda_{\mathbf{x}}}{\lambda_{\mathbf{x}}\boldsymbol{\zeta}(1,1)^\top} & 1\end{pmatrix}\Sigma\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}\lambda_{\mathbf{x}} & \frac{\lambda_{\mathbf{x}}}{\lambda_{\mathbf{x}}\boldsymbol{\zeta}(1,1)^\top} & 1\end{pmatrix} \tag{65}$$

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 (48) to get

$$d\mathbf{x}_t = -\boldsymbol{\Lambda}\mathbf{x}_t dt - d\beta_t + \text{constant}dt \tag{66}$$

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 $\boldsymbol{\Lambda}$.

$$\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = -e^{-\boldsymbol{\Lambda}\tau} \tag{67}$$

where

$$e^{-\boldsymbol{\Lambda}\tau} = \frac{e^{-\nu_1\tau}}{\nu_2 - \nu_1}(\nu_2\mathbf{I} - \boldsymbol{\Lambda}) + \frac{e^{-\nu_2\tau}}{\nu_1 - \nu_2}(\nu_1\mathbf{I} - \boldsymbol{\Lambda}) \tag{68}$$

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

$$\mathbf{P}_t = \boldsymbol{\lambda}_{\boldsymbol{x}}\mathbf{x}_t + \boldsymbol{\lambda}_{\boldsymbol{\beta}}\boldsymbol{\beta}_t + \boldsymbol{\lambda}_V V_t + \bar{\mathbf{P}} \tag{69}$$

$$\frac{\partial \mathbb{E}_t[\mathbf{P}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = \boldsymbol{\lambda}_{\boldsymbol{x}}\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} + \boldsymbol{\lambda}_{\boldsymbol{\beta}} \tag{70}$$

Plug in $\frac{\partial \mathbb{E}_t[\mathbf{x}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top}$ from (67) and $\boldsymbol{\lambda}_{\boldsymbol{\beta}}$ from (58), we have

$$\frac{\partial \mathbb{E}_t[\mathbf{P}_{t+\tau}]}{\partial \boldsymbol{\beta}_t^\top} = -\boldsymbol{\lambda}_{\boldsymbol{x}}e^{-\boldsymbol{\Lambda}\tau} + \boldsymbol{\zeta}^{-1} \tag{71}$$

A.4 Proof of Proposition 3

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

$$Var(dP_{l,t})/dt = \frac{1}{4}(1,1)\Sigma_P \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{4}(1,1)\mathbf{p}^\top\Sigma\mathbf{p} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{72}$$

$$Var(dP_{s,t})/dt = \frac{1}{4}(1,-1)\Sigma_P \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{4}(1,-1)\mathbf{p}^\top\Sigma\mathbf{p} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{73}$$

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

$$\Sigma = \begin{pmatrix} \sigma_\beta^2 & 0 & -\sigma_\beta^2 & 0 & 0 \\ 0 & \sigma_\beta^2 & 0 & -\sigma_\beta^2 & 0 \\ -\sigma_\beta^2 & 0 & \sigma_\beta^2 & 0 & 0 \\ 0 & -\sigma_\beta^2 & 0 & \sigma_\beta^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_V^2 \end{pmatrix} \tag{74}$$

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

$$\lambda_{C,x_C} = \lambda_{S,x_S} \quad \lambda_{C,x_S} = \lambda_{S,x_C} \tag{75}$$

$$\lambda_{C,\beta_C} = \lambda_{S,\beta_S} \quad \lambda_{C,\beta_S} = \lambda_{S,\beta_C} \tag{76}$$

$$\lambda_{C,V} = \lambda_{S,V} \tag{77}$$

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

$$\begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix} \equiv E^\top \boldsymbol{\lambda_x} E, \qquad \begin{pmatrix} l_\beta & 0 \\ 0 & s_\beta \end{pmatrix} \equiv E^\top \boldsymbol{\lambda_\beta} E \tag{78}$$

Hence we can write

$$d \begin{pmatrix} P_{l,t} \\ P_{s,t} \end{pmatrix} = \frac{1}{\sqrt{2}} E^\top \mathbf{p}(\mathbf{s}) = \frac{1}{2} \begin{pmatrix} l_x & l_x \\ s_x & -s_x \end{pmatrix} d\mathbf{x}_t + \frac{1}{2} \begin{pmatrix} l_\beta & l_\beta \\ s_\beta & -s_\beta \end{pmatrix} d\boldsymbol{\beta}_t + \frac{1}{2} \begin{pmatrix} 2 \\ 0 \end{pmatrix} dV_t \tag{79}$$

$$Var(d \begin{pmatrix} P_{l,t} \\ P_{s,t} \end{pmatrix})/dt = \frac{1}{4} \begin{pmatrix} l_x & l_x & l_\beta & l_\beta & 2 \\ s_x & -s_x & s_\beta & -s_\beta & 0 \end{pmatrix} \Sigma \begin{pmatrix} l_x & s_x \\ l_x & -s_x \\ l_\beta & s_\beta \\ l_\beta & -s_\beta \\ 2 & 0 \end{pmatrix} \tag{80}$$

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

$$\Sigma \begin{pmatrix} l_x \\ l_x \\ l_\beta \\ l_\beta \\ 2 \end{pmatrix} = \sigma_\beta^2 (l_x - l_\beta) \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 2\sigma_V^2 \end{pmatrix} \tag{81}$$

Dotting with each row of the left matrix:

$$\text{Row 1:} \quad 2l_x\sigma_\beta^2(l_x - l_\beta) - 2l_\beta\sigma_\beta^2(l_x - l_\beta) + 4\sigma_V^2 = 2\sigma_\beta^2(l_x - l_\beta)^2 + 4\sigma_V^2$$

$$\text{Row 2:} \quad s_x\sigma_\beta^2(l_x - l_\beta) - s_x\sigma_\beta^2(l_x - l_\beta) - s_\beta\sigma_\beta^2(l_x - l_\beta) + s_\beta\sigma_\beta^2(l_x - l_\beta) = 0$$

For column 2:

$$\Sigma \begin{pmatrix} s_x \\ -s_x \\ s_\beta \\ -s_\beta \\ 0 \end{pmatrix} = \sigma_\beta^2 (s_x - s_\beta) \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 0 \end{pmatrix} \tag{82}$$

Dotting with each row:

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

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

$$\text{Var}\left(d \begin{pmatrix} P_{l,t} \\ P_{s,t} \end{pmatrix} \right) / dt = \frac{1}{2} \begin{pmatrix} \sigma_\beta^2 (l_x - l_\beta)^2 + 2\sigma_V^2 & 0 \\ 0 & \sigma_\beta^2 (s_x - s_\beta)^2 \end{pmatrix} \tag{83}$$

Level and spread returns are instantaneously uncorrelated. The variances are

$$\text{Var}(dP_{l,t})/dt = \sigma_\beta^2 (l_x - l_\beta)^2 + 2\sigma_V^2 \tag{84}$$

$$\text{Var}(dP_{s,t})/dt = \sigma_\beta^2 (s_x - s_\beta)^2 \tag{85}$$

To solve for l_x and s_x, we multiply both sides of (55) by E

$$-\begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix} k(I - \tilde{\boldsymbol{\zeta}} \begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix}) = \begin{pmatrix} \psi & 0 \\ 0 & \psi \end{pmatrix} + \gamma \begin{pmatrix} l_x & l_x & l_\beta & l_\beta & 2 \\ s_x & -s_x & s_\beta & -s_\beta & 0 \end{pmatrix} \Sigma \begin{pmatrix} l_x & s_x \\ l_x & -s_x \\ l_\beta & s_\beta \\ l_\beta & -s_\beta \\ 2 & 0 \end{pmatrix} \tag{86}$$

$$-\begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix} k(I - \tilde{\boldsymbol{\zeta}} \begin{pmatrix} l_x & 0 \\ 0 & s_x \end{pmatrix}) = \begin{pmatrix} \psi & 0 \\ 0 & \psi \end{pmatrix} + \gamma \begin{pmatrix} \sigma_\beta^2 (l_x - l_\beta)^2 + 2\sigma_V^2 & 0 \\ 0 & \sigma_\beta^2 (s_x - s_\beta)^2 \end{pmatrix} \tag{87}$$

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

$$\tilde{\boldsymbol{\zeta}} = \begin{pmatrix} \frac{1}{l_\beta} & 0 \\ 0 & \frac{1}{s_\beta} \end{pmatrix} \tag{88}$$

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

$$k\frac{l_x}{l_\beta}(l_x - l_\beta) = \psi + \gamma\left(\sigma_\beta^2(l_x - l_\beta)^2 + 2\sigma_V^2\right) \tag{89}$$

$$k\frac{s_x}{s_\beta}(s_x - s_\beta) = \psi + \gamma\sigma_\beta^2(s_x - s_\beta)^2 \tag{90}$$

define the risk adjusted factors as

$$\phi_l = 1 - \frac{\gamma\sigma_\beta^2 l_\beta}{k} \qquad \phi_s = 1 - \frac{\gamma\sigma_\beta^2 s_\beta}{k} \tag{91}$$

The two equations above can be written as

$$\phi_l(l_\beta - l_x)^2 - l_\beta(l_\beta - l_x) - (\psi + 2\gamma\sigma_V^2)\frac{l_\beta}{k} = 0 \tag{92}$$

$$\phi_s(s_\beta - s_x)^2 - s_\beta(s_\beta - s_x) - \psi\frac{s_\beta}{k} = 0 \tag{93}$$

Hence

$$l_\beta - l_x = \frac{l_\beta(1 + \sqrt{1 + \frac{4\phi_l}{kl_\beta}(\psi + 2\gamma\sigma_V^2)})}{2\phi_l} \tag{94}$$

$$s_\beta - s_x = \frac{s_\beta(1 + \sqrt{1 + \frac{4\phi_s}{ks_\beta}\psi})}{2\phi_s} \tag{95}$$

To see how ψ affects the volatilities, define the instantaneous eigen-portfolio variances

$$V_s \equiv \frac{\mathrm{Var}(dP_{s,t})}{dt} = \sigma_\beta^2(s_\beta - s_x)^2, \tag{96}$$

$$V_l \equiv \frac{\mathrm{Var}(dP_{l,t})}{dt} = \sigma_\beta^2(l_\beta - l_x)^2 + 2\sigma_V^2. \tag{97}$$

the ψ-derivatives are

$$\frac{\partial V_s}{\partial \psi} = 2\sigma_\beta^2(s_\beta - s_x)\,\frac{\partial(s_\beta - s_x)}{\partial \psi} \tag{98}$$

$$\frac{\partial V_l}{\partial \psi} = 2\sigma_\beta^2(l_\beta - l_x)\,\frac{\partial(l_\beta - l_x)}{\partial \psi} \tag{99}$$

From (94) and (95), it is straighforward to see that

$$\frac{\partial(s_\beta - s_x)}{\partial \psi} > 0 \qquad \frac{\partial(l_\beta - l_x)}{\partial \psi} > 0 \tag{100}$$

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

Similarly

$$\frac{\partial V_s}{\partial \gamma} = 2\sigma_\beta^2(s_\beta - s_x)\,\frac{\partial(s_\beta - s_x)}{\partial \gamma} \tag{101}$$

$$\frac{\partial V_l}{\partial \gamma} = 2\sigma_\beta^2(l_\beta - l_x)\,\frac{\partial(l_\beta - l_x)}{\partial \gamma} \tag{102}$$

Plug in $\frac{\partial(s_\beta - s_x)}{\partial \gamma}$ and $\frac{\partial(l_\beta - l_x)}{\partial \gamma}$ we have

$$\frac{\partial V_s}{\partial \gamma} = \frac{2\sigma_\beta^4(s_\beta - s_x)^3}{k\sqrt{1 + \frac{4\phi_s}{k s_\beta}\psi}} > 0 \tag{103}$$

$$\frac{\partial V_l}{\partial \gamma} = \frac{2\sigma_\beta^2(l_\beta - l_x)(\sigma_\beta^2(l_\beta - l_x)^2 + 2\sigma_V^2)}{k\sqrt{1 + \frac{4\phi_l}{k l_\beta}(\psi + 2\gamma\sigma_V^2)}} > 0 \tag{104}$$

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

$$R \equiv \frac{V_s}{V_l} = \frac{\sigma_\beta^2(s_\beta - s_x)^2}{\sigma_\beta^2(l_\beta - l_x)^2 + 2\sigma_V^2} \tag{105}$$

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

$$\frac{\partial \Delta}{\partial \tilde{\psi}} = \frac{1}{kD(a, \tilde{\psi})} > 0, \qquad \frac{\partial \Delta}{\partial a} = -\frac{\phi'(a)\Delta^2 - \Delta - \tilde{\psi}/k}{2\phi(a)\Delta - a} > 0, \tag{106}$$

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 ($\boldsymbol{\zeta}_{12} < 0$), $l_\beta > s_\beta$ 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 (96)–(97),

$$R = \frac{\sigma_\beta^2 \Delta_s^2}{\sigma_\beta^2 \Delta_l^2 + 2\sigma_V^2} < \frac{\sigma_\beta^2 \Delta_l^2}{\sigma_\beta^2 \Delta_l^2 + 2\sigma_V^2} < 1, \tag{107}$$

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

Define $Q \equiv \frac{s_\beta - s_x}{l_\beta - l_x}$, we can write

$$R = \frac{Q^2}{1 + \frac{2\sigma_V^2}{\sigma_\beta^2 (l_\beta - l_x)^2}} \tag{108}$$

Taking log and derivative with respect to ψ we have

$$\frac{\partial \log R}{\partial \psi} = 2\frac{\partial \log Q}{\partial \psi} + \frac{4\sigma_V^2}{\sigma_\beta^2 (l_\beta - l_x)^2 + 2\sigma_V^2} \underbrace{\frac{\partial \log(l_\beta - l_x)}{\partial \psi}}_{>0} \tag{109}$$

Furthermore,

$$\frac{\partial \log Q}{\partial \psi} = \frac{1}{s_\beta - s_x}\frac{\partial (s_\beta - s_x)}{\partial \psi} - \frac{1}{l_\beta - l_x}\frac{\partial (l_\beta - l_x)}{\partial \psi} \tag{110}$$

where

$$\frac{\partial(s_\beta - s_x)}{\partial\psi} = \frac{1}{k\sqrt{1+\frac{4\phi_s}{ks_\beta}\psi}} \tag{111}$$

$$\frac{\partial(l_\beta - l_x)}{\partial\psi} = \frac{1}{k\sqrt{1+\frac{4\phi_l}{kl_\beta}(\psi+2\gamma\sigma_V^2)}} \tag{112}$$

$$\frac{1}{s_\beta - s_x}\frac{\partial(s_\beta - s_x)}{\partial\psi} = \frac{1}{k\sqrt{1+\frac{4\phi_s}{ks_\beta}\psi}}\frac{2\phi_s}{s_\beta(1+\sqrt{1+\frac{4\phi_s}{ks_\beta}\psi})} \tag{113}$$

$$\frac{1}{l_\beta - l_x}\frac{\partial(l_\beta - l_x)}{\partial\psi} = \frac{1}{k\sqrt{1+\frac{4\phi_l}{kl_\beta}(\psi+2\gamma\sigma_V^2)}}\frac{2\phi_l}{l_\beta(1+\sqrt{1+\frac{4\phi_l}{kl_\beta}(\psi+2\gamma\sigma_V^2)})} \tag{114}$$

Define

$$g(a,\tilde{\psi}) \equiv \frac{\phi(a)}{aD(a,\tilde{\psi})(1+D(a,\tilde{\psi}))}, \qquad D(a,\tilde{\psi}) \equiv \sqrt{1+\frac{4\phi(a)}{ka}\tilde{\psi}} \quad \phi(a) \equiv 1 - \frac{\gamma\sigma_\beta^2 a}{k} \tag{115}$$

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 $l_\beta > s_\beta$, we have $g(l_\beta, \psi+2\gamma\sigma_V^2) < g(s_\beta,\psi)$, which implies $\frac{1}{s_\beta - s_x}\frac{\partial(s_\beta - s_x)}{\partial\psi} > \frac{1}{l_\beta - l_x}\frac{\partial(l_\beta - l_x)}{\partial\psi}$. Hence $\frac{\partial \log Q}{\partial\psi} > 0$.

Similarly, taking derivative with respect to γ we have

$$\frac{\partial \log R}{\partial\gamma} = \frac{\partial \log V_s}{\partial\gamma} - \frac{\partial \log V_l}{\partial\gamma} \tag{116}$$

where

$$\frac{\partial \log V_s}{\partial\gamma} = \frac{1}{V_s}\frac{\partial V_s}{\partial\gamma} = \frac{2\sigma_\beta^2(s_\beta - s_x)}{k\sqrt{1+\frac{4\phi_s}{ks_\beta}\psi}} \tag{117}$$

$$\frac{\partial \log V_l}{\partial\gamma} = \frac{1}{V_l}\frac{\partial V_l}{\partial\gamma} = \frac{2\sigma_\beta^2(l_\beta - l_x)}{k\sqrt{1+\frac{4\phi_l}{kl_\beta}(\psi+2\gamma\sigma_V^2)}} \tag{118}$$

Hence

$$\frac{\partial \log R}{\partial \gamma} = \frac{2\sigma_\beta^2}{k}\left(\frac{s_\beta - s_x}{\sqrt{1+\frac{4\phi_s}{ks_\beta}\psi}} - \frac{l_\beta - l_x}{\sqrt{1+\frac{4\phi_l}{kl_\beta}(\psi+2\gamma\sigma_V^2)}}\right) \tag{119}$$

$$= \frac{2\sigma_\beta^2}{k}\left(\frac{s_\beta}{2\phi_s}\left(1+\frac{1}{\sqrt{1+\frac{4\phi_s}{ks_\beta}\psi}}\right) - \frac{l_\beta}{2\phi_l}\left(1+\frac{1}{\sqrt{1+\frac{4\phi_l}{kl_\beta}(\psi+2\gamma\sigma_V^2)}}\right)\right) \tag{120}$$

we have

$$\frac{s_\beta}{2\phi_s}\left(1+\frac{1}{\sqrt{1+\frac{4\phi_s}{ks_\beta}\psi}}\right) < \frac{s_\beta}{\phi_s} \tag{121}$$

and

$$\frac{l_\beta}{2\phi_l}\left(1+\frac{1}{\sqrt{1+\frac{4\phi_l}{kl_\beta}(\psi+2\gamma\sigma_V^2)}}\right) > \frac{l_\beta}{2\phi_l} \tag{122}$$

Under the condition that $\frac{2s_\beta}{\phi_s} < \frac{l_\beta}{\phi_l}$, 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

$$V_l = \frac{1}{4}\mathrm{Var}(dP_{C,t} + dP_{S,t}) = \frac{\sigma_P^2(1+\rho)}{2}, \quad V_s = \frac{1}{4}\mathrm{Var}(dP_{S,t} - dP_{C,t}) = \frac{\sigma_P^2(1-\rho)}{2} \tag{123}$$

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

$$R \equiv \frac{V_s}{V_l} = \frac{1-\rho}{1+\rho} \quad \Longleftrightarrow \quad \rho = \frac{1-R}{1+R} \tag{124}$$

Since ρ is strictly decreasing in R,

$$\frac{\partial \rho}{\partial \theta} = -\frac{2}{(1+R)^2}\frac{\partial R}{\partial \theta} \tag{125}$$

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,

$$\theta_i(\tau) = \tilde{d}_{0,i} + \tilde{d}_{1,i}\, e^{-\nu_1 \tau} + \tilde{d}_{2,i}\, e^{-\nu_2 \tau}, \quad i = l, s, \tag{126}$$

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

$$\hat{m}_i = -\left(\nu_1\, \tilde{d}_{1,i} + \nu_2\, \tilde{d}_{2,i}\right). \tag{127}$$

We estimate the parameters in (126) using nonlinear least squares (NLS) applied to the local projection specification (41):

$$\Delta_\tau y_{i,t} = a_i + \left(\tilde{d}_{0,i} + \tilde{d}_{1,i}\, e^{-\nu_1 \tau} + \tilde{d}_{2,i}\, e^{-\nu_2 \tau}\right) \beta_{i,t} + c_{i,\tau}\, \Delta_\tau y_{i,t-\tau-1} + \varepsilon_{i,t+\tau}. \tag{128}$$

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:

$$\begin{aligned} \Delta_\tau y_{i,t} &= a_i^y + c_{i,\tau}^y\, \Delta_\tau y_{i,t-\tau-1} + \varepsilon_{i,t+\tau}^y, \\ \beta_{i,t} &= a_i^\beta + c_{i,\tau}^\beta\, \Delta_\tau y_{i,t-\tau-1} + \varepsilon_{i,t+\tau}^\beta, \end{aligned} \tag{129}$$

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:

$$\Delta_\tau \check{y}_{i,t} = \left(\tilde{d}_{0,i} + \tilde{d}_{1,i}\, e^{-\nu_1 \tau} + \tilde{d}_{2,i}\, e^{-\nu_2 \tau}\right) \check{\beta}_{i,t} + \varepsilon_{i,t+\tau}. \tag{130}$$

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

B.2 Support Identification Algorithm

This section describes the data-driven procedure for identifying the estimation window $[\underline{\tau}, \overline{\tau}]$ over which we fit the decay function (referenced in Section 3.5).

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 window $[\underline{\tau}, \overline{\tau}] = [0, 10]$ 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: Zero-crossing of the IRF. Check whether the fitted polynomial $\hat{f}(\tau)$ itself crosses zero at any point after the start of the support. If so, set the end point $\overline{\tau}$ to the first such zero-crossing.

Step 5: Window selection. Set the start of the support as the point with the most extreme response among $\tau = 0$ and all peak candidates:

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

If the end point was not determined in Step 4, set it as the most extreme trough:

$$\overline{\tau} = \arg\min \left\{ |\hat{f}(\tau_{\text{trough},1}^*)|, \ldots, |\hat{f}(\tau_{\text{trough},K}^*)| \right\}.$$

B.3 Constrained Estimation and Bootstrap Inference

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

$$\frac{m_s}{m_l} \in \left[\frac{\sigma_{1,s}}{\sigma_1^2},\, 1\right]. \tag{131}$$

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 (131) 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 check whether the unconstrained OLS estimates of m_l and m_s satisfy the theoretical bounds. If they do, we report the unconstrained estimates with date-clustered standard errors. If the bounds are violated, we solve the constrained least squares problem

$$\min_{\mathbf{b}} \frac{1}{2}\mathbf{b}^\top(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})\mathbf{b} - (\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}})^\top\mathbf{b} \quad \text{subject to} \quad \frac{m_s}{m_l} \in \left[\frac{\sigma_{1,s}}{\sigma_1^2}, 1\right], \tag{132}$$

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. When the constraint binds, standard analytical standard errors are not valid. We use a date-cluster bootstrap: 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 $\hat{m}_l$, $\hat{m}_s$, $\hat{C}_{r,l}$, and $\hat{C}_{r,s}$ are computed as the standard deviation of the bootstrap distribution.

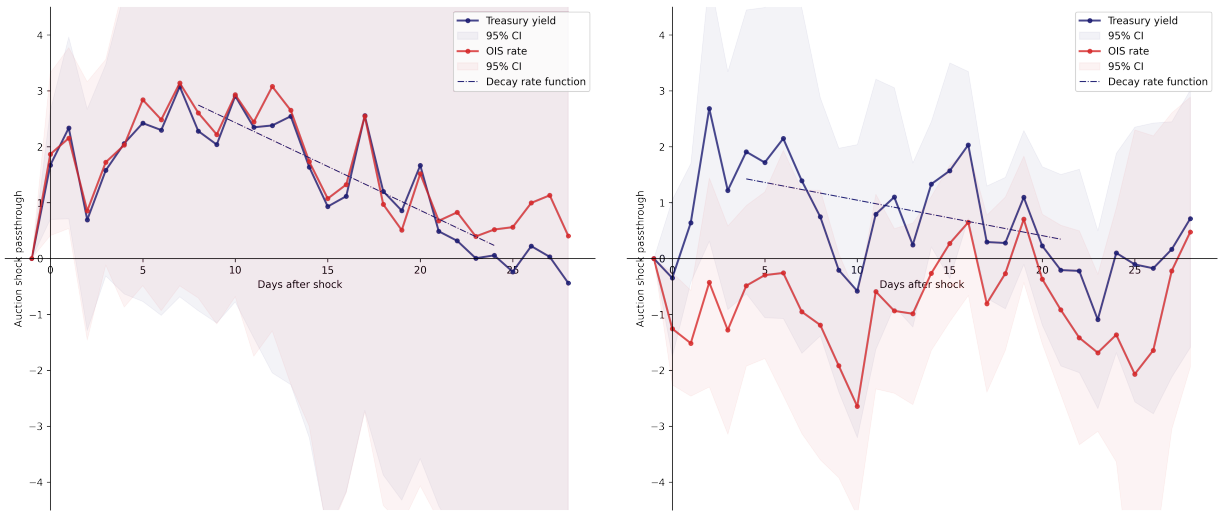
R^2 loss diagnostic. To assess the cost of imposing the theoretical bounds, we report the R^2 loss: the difference in R^2 between the unconstrained and constrained fits. A small R^2 loss indicates that the constraint is nearly satisfied by the data and the constrained estimates are close to the unrestricted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

C Crisis Episode Impulse Responses: Levels

Figure 7 shows the impulse responses of Treasury yields and the OIS rate to 5-year Treasury note auction demand shocks during the two crisis episodes.

Figure 7: 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).