Demand-and-Supply Imbalance Risk and Long-Term Swap Spreads*

Samuel G. Hanson
Harvard Business School

Aytek Malkhozov
Queen Mary University of London

Gyuri Venter
Warwick Business School

September 13, 2023

Abstract

We develop and test a model in which swap spreads are determined by end users' demand for and constrained intermediaries' supply of long-term interest rate swaps. Swap spreads reflect compensation both for using scarce intermediary capital and for bearing convergence risk—i.e., the risk spreads will widen due to a future demand-and-supply imbalance. We show that a proxy for the intermediated quantity of swaps—dealers' net position in Treasuries—flipped sign during the Global Financial Crisis when swap spreads turned negative and that this variable predicts the excess returns on swap spread trades. Exploiting our model's sign restrictions, we identify shifts in demand and supply and find that both contribute significantly to the volatility of swap spreads.

Keywords: Swap spreads, limits to arbitrage, intermediary capital constraints.

JEL Classification: G12, E43

*Emails: shanson@hbs.edu, a.malkhozov@qmul.ac.uk, and gyuri.venter@wbs.ac.uk. We thank Benjamin Hébert, Urban Jermann, Andrei Kirilenko, Sven Klingler, Yoshio Nozawa, Ajay Subramanian, Fabrice Tourre, Dimitri Vayanos, Paul Whelan, and participants at numerous seminars and conferences for helpful comments. We are grateful to Chris Anderson, Zhiguo He, and Sven Klingler for sharing their data.

1 Introduction

An interest rate swap is an agreement to exchange a series of fixed interest payments for a series of floating interest payments based on the future realizations of a short-term reference rate. End users, including both financial institutions and non-financial firms, use these derivatives to manage their exposure to interest-rate risk. With an estimated $115 trillion of outstanding notional value as of 2020, the market for interest rate swaps is the largest derivatives market in the U.S.

From their advent in the 1980s until 2008, the fixed rates on swaps had always remained above the corresponding rates on like-maturity government bonds. In other words, swap spreads—the difference between swap rates and like-maturity government bond yields—had been uniformly positive. However, beginning in the 2008 Global Financial Crisis (GFC), the fixed rates on long-term (e.g., 30-year) swaps fell below government bond yields, resulting in negative long-term swap spreads. Negative swap spreads seemingly represent a pure arbitrage opportunity, a puzzle in such a large and liquid market (Boyarchenko et al., 2018). Negative swap spreads have been alternately attributed to large increases in end-user demand for long-dated swaps or to rising balance-sheet costs at the financial intermediaries that supply swaps. See Klingler and Sundaresan (2019) and Jermann (2020) for, respectively, a demand and a supply perspective on negative swap spreads. Of course, both demand and supply forces may have contributed to the negative long-term swap spreads observed in recent years. And, disentangling these forces can help us understand whether a given movement in spreads is due to a shift in end-user demand or intermediary supply (or both).

In this paper, we develop a tractable theoretical framework to study the joint effects of end-user demand and intermediary supply on swap spreads. We show how data on prices and intermediated quantities can be used to separately identify end users' net demand for and intermediaries' supply of swaps. We can then decompose the observed variation in swap spreads into the contributions of these two forces. We highlight a key determinant of long-dated swap spreads that has been overlooked in the recent literature: intermediaries need to be compensated for the risk that spreads may temporarily widen because of future shocks to demand or supply. We argue that compensation for such demand-and-supply imbalance risk explains a significant fraction of the returns to swap spread arbitrage. Thus, our paper hearkens back to the earlier literature pioneered by De Long et al. (1990), emphasizing that "convergence risk" is a significant limit to arbitrage.

We begin by building an equilibrium model of the interest rate swap market in the tradition of Vayanos and Vila (2021). In our model, the demand of end users such as pension funds and non-financial firms to receive the fixed rate on long-term swaps is not naturally offset by opposing demands from other end users. Since swaps are in zero net supply, this time-varying net end-user demand must be absorbed in equilibrium by risk-averse and leverage-constrained intermediaries. Motivated by the evidence in Siriwardane et al. (2021), we adopt a "segmented-markets" view and assume that the relevant intermediaries are quite specialized in the swap market. In practice, we associate the intermediaries in our model with swap desks at broker-dealers or swap traders at fixed-income hedge funds. These specialized intermediaries hedge the interest-rate risk arising from their swap positions in the Treasury market. As a result, they are only concerned with the relative valuation of swaps and Treasuries, namely, with the level of swap spreads and the short-rate differential—i.e., the difference between the short-term floating rate referenced by the swap and the short-term financing rate applicable to Treasuries.

In our model, intermediaries' willingness and ability to intermediate between the swap and Treasury markets is limited by two factors. First, intermediaries are risk averse, so the risk that they may suffer short-term, mark-to-market losses limits their willingness to engage in the swap spread trade. Second, intermediaries face a potentially binding balance-sheet constraint that can limit their ability to undertake even riskless trades. Furthermore, the amount of capital intermediaries have to deploy fluctuates over time. Thus, relative to other Vayanos and Vila (2021)-style models where intermediaries must absorb end-user demand shocks, our model adds an intermediary balance-sheet constraint and the possibility of independent intermediary supply shocks.

If the short-rate differential is non-zero, which is the case for swaps whose floating leg is tied to the 3-month London Interbank Offer Rate (LIBOR), a non-zero swap spread is consistent with the absence of arbitrage. Furthermore, when the short-rate differential fluctuates over time, swap spread trades expose intermediaries to a form of fundamental risk. Prior to the 2008 GFC, our model suggests that LIBOR swap spreads were positive because (i) the short-rate differential for LIBOR swaps was always positive and (ii) intermediaries had to accommodate a net demand by end users to pay the fixed swap rate and were "long" swap spreads.

While we allow for a non-zero short-rate differential to speak to the LIBOR swaps that historically dominated the market, we will regularly emphasize the case where the short-rate differential is always zero, which is the case for swaps whose floating leg is tied to the Secured Overnight Financing Rate (SOFR) for Treasuries. A non-zero SOFR swap spread represents a failure of the Law of One Price (LoOP) and cannot survive in our model's stable equilibrium in the absence of intermediary balance-sheet constraints. However, the spreads on both long-dated LIBOR swaps and SOFR swaps turned negative in late 2008 and have remained negative since.

How can we understand the highly negative swap spreads witnessed since 2008? The potential for binding intermediary balance-sheet constraints opens the door to non-zero SOFR swap spreads in as in Gârleanu and Pedersen (2011). When balance-sheet constraints bind, swap spreads are driven by shocks to end-user demand and intermediary wealth, both of which affect the tightness of intermediaries' leverage constraint and hence their ability to supply swaps. Moreover, our model suggests that *convergence risk* is a key determinant of long-dated swap spreads. Arbitraging long-dated swap spreads not only consumes scarce capital, but is also risky for specialized intermediaries: future shocks to either demand or supply may lead spreads to widen in magnitude, triggering short-term losses for intermediaries. Intermediaries will require compensation for bearing such *demand-and-supply imbalance risk*, which increases the magnitude of swap spreads, particularly at longer maturities.[1] The presence of convergence risk further implies that the size of intermediaries' positions in the swap spread trade should predict the returns to this trade, even after controlling for intermediaries' current balance sheet costs.

As discussed in the Treasury Borrowing Advisory Committee's recent report on the swap market (TBAC, 2021) and documented below, there are strong indications that there was a major

[1]This observation is related to the ideas in Du et al. (2022), who argue that asset price fluctuations stemming from time variation in intermediary constraints command a positive risk premium.

regime shift in end-user net demand in late 2008, with net demand swinging from a desire to pay the fixed swap rate to a desire to receive the fixed swap rate. As a result, intermediaries' net position in the swap spread trade and hence the compensation they require for holding this position—both the compensation for consuming scarce capital and that for bearing convergence risk—also likely flipped signs during the GFC. Seen through our model, this shift in end-user demand and intermediaries' net position explains why swap spreads turned negative in late 2008 and have remained negative since. By contrast, while a rise in intermediary balance-sheet costs following the GFC may have raised the absolute magnitude of swap spreads, our model suggests that, a rise in balance-sheet costs by itself cannot explain why swap spreads flipped signs.

Naturally, demand shocks and supply shocks induce different comovement patterns between swap spreads and intermediaries' swap positions in our model. Specifically, positive shocks to end-user demand increase the magnitude of both swap spreads and intermediaries' swap positions in equilibrium. By contrast, positive shocks to intermediaries' supply of swaps—i.e., positive shocks to intermediaries' wealth—decrease the magnitude of swap spreads but increase the magnitude of swap positions. Under certain linearity-generating assumptions, we show that our model admits a linear equilibrium in which swap spreads are an affine function of three exogenous state variables that govern the short-rate differential, the level of end-user demand, and the level of intermediary wealth, respectively. In this case, our model's linear equilibrium has a representation as a structural vector auto-regression (VAR) in which demand and supply shocks can be identified using sign restrictions on the responses of swap spreads and intermediary positions to these shocks. We also show the main qualitative predictions of our model hold even if we do not impose these linearity-generating assumptions, although the model naturally loses its structural VAR representation in this case.

To take our model to the data, we need a proxy for intermediaries' net position in the swap spread trade. Our proxy for intermediaries' net position in the receive-fixed swap spread trade is negative one times primary dealers' net long position in Treasuries ($-1\times$*PD-UST-Net*$_t$). Equivalently, by market clearing, our proxy for end-users' net position in receive-fixed swaps is *PD-UST-Net*$_t$. The intuition is as follows. Dealers are critical intermediaries in the swap market (TBAC, 2021). However, they seek to have minimal net exposure to changes in the overall level of long-term rates. As a result, dealers' net position in Treasuries—especially in Treasuries maturing in more than one year—is a hedge that partially mirrors their net position in interest rate swaps

Admittedly, this proxy is far from perfect; a variety of factors other than the scale of their net position in swaps affect dealers' net Treasury position. For instance, dealers play a role in intermediating Treasury auctions, absorbing some portion of new issuance into short-term inventory and distributing these securities to end investors over time (Fleming and Rosenberg, 2008). Furthermore, dealers use Treasuries to hedge their inventories of a variety of other fixed-income instruments beyond swaps—e.g., corporate bonds. We discuss these issues in greater detail below, but our overall conclusion is that *PD-UST-Net*$_t$ is a useful, albeit noisy, proxy for the scale of end-users' net position. Since we use *PD-UST-Net*$_t$ as an independent variable in our regression specifications, this measurement error will attenuate our findings, biasing us against finding a significant relationship between *PD-UST-Net*$_t$ and swap spreads or swap spread trade returns.

To validate our use of $PD\text{-}UST\text{-}Net_t$ as proxy for end-users' net position in receive-fixed swaps, we begin by showing that long-dated swap spreads and $PD\text{-}UST\text{-}Net_t$ move inversely over time. Furthermore, consistent with the regime change in end-user demand referenced above, both long-dated swap spreads and $PD\text{-}UST\text{-}Net_t$ simultaneously flipped signs in early 2009. Specifically, $PD\text{-}UST\text{-}Net_t$ moved from negative to positive in early 2009 at the same time that swap spreads went from positive to negative; see Panel A of Figure 1.

Focusing on the post-GFC data, we next show that $PD\text{-}UST\text{-}Net_t$ negatively predicts the future returns on long-dated swap spread positions, even controlling for balance-sheet costs. Specifically, when $PD\text{-}UST\text{-}Net_t$ is higher, the future returns on long-dated swap spread positions tend to be lower on average. Interpreted through the lens of our model, the idea is that a higher level of $PD\text{-}UST\text{-}Net_t$ signals that end-users have a larger net position in receive-fixed swaps. To induce risk averse intermediaries to take the short side of this trade—i.e., to take a large net position in pay-fixed swap spread trade, the expected returns on the receive-fixed swap spread trade must decline. We show that this predictability is strongest when we control for other factors—e.g., the gross quantity of Treasury issuance—that are unrelated to the scale of dealers' swap spread positions but that independently affect $PD\text{-}UST\text{-}Net_t$.

Having validated $PD\text{-}UST\text{-}Net_t$ as a useful proxy for end-users' net position, we estimate the structural VAR implied by our linear model during the post-GFC period. Specifically, we set-identify the parameters of our structural VAR using the sign restrictions approach of Uhlig (2005). Using these VAR estimates, we then extract the latent underlying demand and supply factors from swap spreads and $PD\text{-}UST\text{-}Net_t$ and decompose swap spreads into the respective contributions of these factors. Looking over our entire post-GFC sample, we find that demand and supply shocks play roughly equal roles in explaining the time-series variation in long-term swap spreads. At the same time, our decomposition sheds light on whether a given historical movement in spreads was due to a shift in demand, in supply, or in both. For instance, our decomposition suggests that a large inward shift in intermediary supply from late-2014 until late-2015 pushed swap spreads far into negative territory. This estimated inward supply shift coincides with a series of regulatory changes that arguably increased the balance-sheet costs faced by intermediaries, including the finalization of the Supplementary Leverage Ratio in September 2014 and the implementation of the Volcker Rule in July 2015 (Boyarchenko et al., 2020). However, the period of highly negative swap spreads between late 2014 and 2018 cannot be solely attributed to the shift in intermediary supply. Indeed, from mid-2016, our estimates suggest that rising end-user demand to receive the fixed rate pushed swap spreads even further below zero.

Going further, this decomposition allows us to shed light on the primitive factors that drive swap spreads by examining the time-series correlates of our estimated demand and supply factors. Echoing Feldhütter and Lando (2008) and Hanson (2014), our analysis suggests that hedging demand from mortgage investors is the most significant driver of time variation in net end-user demand for swaps. However, consistent with Klingler and Sundaresan (2019), asset-liability management by pension funds also plays a role in shaping the demand for swaps. We also find more modest evidence that intermediary supply is smaller when proxies for intermediary balance-sheet capacity are depressed.

Finally, we examine the respective roles of end-user demand and intermediary supply in shaping the expected returns to swap arbitrage. While demand and supply both play important roles in driving the variation in swap spreads, we should expect end-user demand to be a stronger predictor of the returns to swap arbitrage than supply. Intuitively, positive demand shocks simultaneously increase the compensation intermediaries require for committing their scarce balance-sheet capacity to swap arbitrage as well as the required compensation for bearing swap spread risk. By contrast, a negative supply shock—i.e., a negative shock to intermediaries' wealth—reduces intermediaries' exposure to swap spread risk and the associated risk premium while increasing the required compensation for using scarce balance sheet. Because these two effects partially offset each other, supply shocks have a smaller impact on the returns to swap arbitrage than demand shocks. We find some evidence of this asymmetry in the data, providing further support for the demand-and-supply imbalance risk channel.

While our model and empirical analysis focus on the interest rate swap market, our insights are more general and our model could be readily applied to a host of other long-dated near-arbitrage spreads, including deviations from covered interest rate parity (CIP) in foreign exchange markets and the CDS-bond basis in credit markets. Indeed, we believe that our model—an intermediary-based pricing model with separate shocks to end-user demand and intermediary supply—is a useful contribution in its own right. This is because models with demand effects generally abstract away from intermediary wealth effects—i.e., supply shocks.

Our model builds on Vayanos and Vila (2021) and, as a result, is related to the growing literature on demand factors in the government bond market; see Greenwood and Vayanos (2014), Hanson (2014), Malkhozov et al. (2016), Haddad and Sraer (2020), Gourinchas et al. (2020), and Greenwood et al. (2022), among others. We depart from this prior literature along two dimensions. First, we consider long-maturity swap spreads rather than long-maturity bond yields. Second, we allow for the variation in both end-user demand and intermediary supply in the swap market—in this way, our work is related to the model of the government bond market developed in Kekre et al. (2022)—and use our model to empirically disentangle these two forces.

Our work is also related to De Long et al. (1990), who show that noise trader demand shocks can create a form of convergence risk that deters rational arbitrageurs from aggressively betting against a LoOP violation. In their model, an equilibrium where the LoOP fails and arbitrage is limited by convergence risk exists alongside a more standard equilibrium in which arbitrageurs enforce the LoOP. By contrast, we show that LoOP violations do not have to rely on the unstable type of equilibrium considered in De Long et al. (1990). LoOP violations arise in our model's unique stable equilibrium because intermediaries are subject to a binding balance-sheet constraint as in Gârleanu and Pedersen (2011). Once these primitive LoOP violations arise, they are amplified by convergence risk arising from demand and supply shocks.

Swap rates and Treasury yields have been extensively studied in the previous literature. One strand of this literature calibrates dynamic term structure models to understand the dynamics of swap spreads; see, for instance, Duffie and Singleton (1997), Lang et al. (1998), Collin-Dufresne and Solnik (2001), Liu et al. (2006), Feldhütter and Lando (2008), and Augustin et al. (2021), among others. Two recent papers that are important precursors to our study are Klingler and

Sundaresan (2019) and Jermann (2020) who focus on, respectively, the role of end-user demand and constrained intermediary supply in explaining the emergence of negative swap spreads during the GFC. Klingler and Sundaresan (2019) argue that changes in pension underfunding lead to shifts in the demand to receive the fixed swap rate and that the rise in aggregate underfunding during the GFC helps explain the emergence of negative swap spreads. We find evidence consistent with this channel. Jermann (2020) emphasizes the role of constrained intermediary supply and argues that equilibrium swap spreads can be negative even if intermediaries do not have to absorb any net end-user demand in the swap market.[2] Our emphasis on intermediary balance sheet constraints builds heavily on Jermann (2020). Relative to these important precursors, however, our framework allows us to highlight, both theoretically and empirically, how shifts in *both* end-user demand and intermediary supply contribute to movements in swap spreads and the negative swap spreads observed since the GFC.

In contemporaneous work, Du et al. (2022) also draw a connection between the regime shift in swap spreads and the regime shift in dealers' net Treasury position after the 2008 Global Financial Crisis. Taking swap yields as given, they use an affine term structure model to compute the yields at which a leverage-constrained dealer should be willing to go net short or net long Treasuries: the net-short yield is below the swap yield and the net-long yield is above. They note that Treasuries yields moved from the net-short yield they compute to their net-long yield in late 2008 and early 2009—i.e., that swap spreads changed sign—at the same time $PD\text{-}UST\text{-}Net_t$ moved from negative to positive. We believe the two papers are complementary. Our focus is on disentangling the role of end-user demand and intermediary supply shocks in shaping the level of swap spreads and documenting the importance of demand-and-supply imbalance risk. By contrast, Du et al. (2022) are primarily interested in understanding the regime shift that occurred during the GFC.

Finally, our work is also related to Cohen et al. (2007), Chen et al. (2018), Goldberg and Nozawa (2021), and Fontaine et al. (2023), who identify demand and supply shifts in shorting, index option, and corporate bond markets, and intermediary leverage, respectively. We also build on Cieslak and Pang (2021), who use sign restrictions to identify latent factors that drive time-variation in asset prices.

The rest of the paper is organized as follows. Section 2 provides background on the interest rate swap market. Section 3 presents a theoretical framework to study the impact of end-user demand and intermediary supply on swap spreads and derives testable predictions. Section 4 describes our data and presents our main empirical results. Section 5 concludes. An Internet Appendix presents a range of additional theoretical and empirical results.

2 Background on interest rate swaps

This section provides the necessary background on the interest rate swap market and can be skipped by readers who are well-versed in the swap literature. After explaining the mechanics

[2] Jermann (2020) assumes intermediaries want to hold long-term Treasuries to capture a positive term premium, only intermediaries can arbitrage between Treasuries and swaps, and intermediaries face greater balance sheet costs from taking positions in Treasuries than swaps. Under these assumptions, equilibrium swap spreads can be negative even if there is zero net end-user demand for swaps and intermediaries have zero net exposure to swap spreads.

of these derivative contracts, we discuss the participants in the swap market, dividing them into "end users" and specialized "intermediaries." We then review the no-arbitrage logic that makes both non-zero SOFR swap spreads and negative LIBOR swap spreads so surprising. Finally, we describe the major changes this market has undergone since the 2008 Global Financial Crisis.

2.1 Swap mechanics

An interest rate swap is an agreement between two counterparties to exchange a series of interest payments over time based on some notional principal amount. In a plain-vanilla swap, the first counterparty pays the second a series of pre-determined payments based on the fixed swap rate set at the contract's inception; the second counterparty pays the first a series of floating and initially unknown payments based on the future realizations of some short-term reference rate. As a result, the counterparty who is receiving (paying) the fixed rate acquires a financial exposure similar to that obtained by borrowing (investing) cash at the short-term interest rate and taking a long (short) position in long-term bonds. The fixed swap rate is set so that the swap has zero value at inception. Thus, the fixed swap rate is akin to a par coupon yield derived from the underlying reference curve. A par swap spread is the difference between the fixed swap rate and the par coupon yield of a Treasury bond with the same maturity.

Historically, the floating leg on most swaps was tied to the 3-month London Interbank Offer Rate. LIBOR is an indicative, unsecured borrowing rate for major global banks and the 3-month LIBOR embeds some small amount of credit risk. In recent years, swaps tied to overnight unsecured interbank borrowing rates and overnight secured borrowing rates—which embed virtually no credit risk—have been replacing LIBOR-based swaps and LIBOR was discontinued at the end of 2021. Specifically, Overnight Index Swaps (OIS) are tied to the effective federal funds rates (the volume-weighted median rate on overnight unsecured interbank loans) and SOFR swaps are tied to the Secured Overnight Financing Rate (the volume-weighted median rate in the market for overnight repurchase agreements backed by Treasuries).

The swap market is the largest U.S. derivatives market with over \$115 trillion of outstanding notional value and an estimated gross market value of \$2.3 trillion in 2020.[3] The market is extremely liquid with an average daily volume of roughly \$500 billion in 2020.[4] Based on Bloomberg data, the typical bid-ask yield spread for 10-year swaps is 0.5 basis points whereas the typical bid-ask spread for the on-the-run 10-year Treasury note is 0.25 basis points.

2.2 Swap market participants

It is useful to group swap market participants into "end users" and specialized "intermediaries." End users use swaps to manage their pre-existing exposures to interest-rate risk. To clear the market, specialized intermediaries—primarily swap desks at broker-dealers and swap traders at fixed-income hedge funds—must accommodate the net end-user demand to either receive or pay

[3]See https://stats.bis.org/statx/srs/table/d5.1. Gross market value is the absolute value of the market value of all outstanding receive-fixed swaps. There were \$21 trillion of outstanding U.S. Treasures at 2020 year-end.

[4]See https://www.clarusft.com/2020-ccp-volumes-and-market-share-in-ird/. By way of comparison, SIFMA reports that the average daily volume in Treasuries was roughly \$600 billion in 2020.

the fixed swap rate. These intermediaries hedge the interest-rate risk associated with their swap positions in the Treasury market (TBAC, 2021). As a result, intermediaries are only concerned with the relative valuation of swaps and Treasuries—i.e., with the level of swap spreads and any differential between the short-term rate referenced by swaps and the short-term Treasury financing rate. In addition to the P&L on their swap spread positions, intermediaries must also weigh the fact that these positions consume scarce risk-bearing capital.

Swap end users are agents who want long or short exposure to long-term bonds, but who for regulatory, accounting, or other frictional reasons prefer to obtain this exposure using swaps rather than Treasuries. As explained in the Treasury Borrowing Advisory Committee's report on the swap market (TBAC, 2021), there are several important groups of end users:

- **Insurers and pensions** use swaps to manage their exposure to interest-rate risk. Insurers and pensions typically receive the fixed swap rate on net—using swaps to add duration—because the duration of their liabilities exceeds that of their on-balance sheet assets. Klingler and Sundaresan (2019) argue that pensions' desire to receive fixed increases when they become more underfunded. Furthermore, since the convexity of their liabilities exceeds that of their assets, insurers and pensions tend to enter into additional receive-fixed swaps when long-term rates fall to dynamically manage their interest-rate exposure (Domanski et al., 2017).
- **Commercial banks** typically receive fixed on net (Begenau et al., 2020). Although banks borrow short-term and lend long-term, banks' net income is typically hurt by declining interest rates because their loans reprice more quickly than their deposits (Driscoll and Judson, 2013; Drechsler et al., 2021). To offset these exposures, banks generally receive fixed.
- **Non-financial corporations** typically receive fixed on net to convert fixed-rate bond issues into synthetic floating-rate funding.
- **Relative-value mortgage investors** pay the fixed swap rate on net. These investors—most prominently Fannie Mae and Freddie Mac—attempt to exploit the fact that pass-through mortgage-backed securities (MBS) sometimes trade cheap relative to a dynamic replicating portfolio. Mortgage investors prefer to hedge MBS with swaps instead of Treasuries for regulatory and accounting reasons and because swaps have historically been a more effective hedge. When long-term rates fall, expected mortgage prepayments rise, causing MBS duration to decline (i.e., MBS have negative convexity). This prompts mortgage investors to enter receive-fixed swaps to reduce the size of their pay-fixed hedge positions (Perli and Sack, 2003; Feldhütter and Lando, 2008; Hanson, 2014; Malkhozov et al., 2016).
- **Mortgage servicers** are institutions that earn a stream of fees to process mortgage payments; they collect monthly mortgage-related payments from homeowners and pass them along to MBS investors, local tax authorities, and property insurers. Mortgage servicers have an exposure similar to the holder of an interest-only (IO) MBS strip, which typically has a *negative* duration. To offset the negative duration of their assets, servicers are typically net fixed receivers. Further, since IO strips are negatively convex just like pass-through MBS, servicers tend to increase their receive-fixed positions when rates fall.

- **Fixed-income money managers** use swaps to adjust the duration of their portfolios and are generally net fixed payers. These investors typically hold cash bonds and prefer to use pay-fixed swap positions to attain their duration targets.

In summary, most major groups of end users are typically net fixed receivers. The main net fixed payers are relative-value mortgage investors and other money managers. Furthermore, due to convexity-driven hedging, net end-user demand to receive fixed typically rises when rates decline.

2.3 No-arbitrage pricing of swaps

To begin, note that a non-zero SOFR swap spread constitutes a failure of the Law of One Price and implies the existence of a feasible, zero-cost portfolio that generates a riskless stream of positive cashflows at all dates prior to maturity. This is because a SOFR swap's floating leg is tied to the overnight SOFR reference rate published by the Federal Bank of New York and *perfectly* tracks the short-term rate at which intermediaries can finance their Treasury positions.[5] Specifically, if the SOFR swap spread is positive, entering a receive-fixed SOFR swap and taking an offsetting short position in like-maturity Treasuries (and investing the proceeds at SOFR) generates a stream of riskless cashflows equal to the swap spread. Conversely, if the SOFR swap spread is negative, entering a pay-fixed SOFR swap and taking a long position in Treasuries (financed at SOFR) generates a stream of riskless cashflows equal to the negative one times the swap spread. Thus, absent frictions, no arbitrage logic implies that SOFR swap spreads must be zero.

By contrast, since the 3-month LIBOR exceeds the secured financing rate for Treasuries, a positive spread on a LIBOR swap does not represent a failure of no arbitrage. This is because the cashflows from receiving fixed on a LIBOR swap and shorting Treasuries can be negative if the LIBOR minus SOFR short-rate differential is large enough. However, assuming that LIBOR exceeds SOFR in all possible states, a negative LIBOR swap spread is a violation of no arbitrage. In this case, paying fixed on a LIBOR swap and taking a long position in Treasuries is a zero-cost portfolio with strictly positive cashflows in all possible states.

In sum, the negative LIBOR swap spreads and non-zero SOFR swap spreads witnessed since 2008 are inconsistent with frictionless, no-arbitrage pricing. Thus, it makes sense to build a model where intermediaries face constraints that prevent them from eliminating these opportunities.

2.4 The evolution of the swap market

From the inception of the swap market in the 1980s until the height of the Global Financial Crisis (GFC) in late 2008, LIBOR-based swap spreads had always been positive—i.e., swap yields had always exceeded like-maturity Treasury yields. The most straightforward explanation for positive swap spreads is that LIBOR always exceeds Treasury repo rates because (i) LIBOR is an unsecured 3-month bank borrowing rate that includes compensation for credit risk (Collin-Dufresne and

[5]Technically, SOFR swaps deviate ever so slightly from this theoretical limit: due to frictions in the repo market, individual intermediaries typically finance their Treasury holdings at rates that differ slightly from SOFR (the volume-weighted median financing rate). However, spreads within the Treasury repo market are very small—an order of magnitude smaller than the spread between LIBOR and repo rates—and appear to be quite stable, so we ignore this detail in the remainder of the paper.

Solnik, 2001) and (ii) Treasury yields and repo rates are depressed by a money-like convenience premium that is specific to these extremely safe and liquid assets (Feldhütter and Lando, 2008; Krishnamurthy and Vissing-Jorgensen, 2012).

However, there is evidence that demand-and-supply forces played a role in supporting the level of swap spreads already before 2008. Pre-GFC, hedged mortgage investors played a dominant role in the swap market and there was generally net demand from end users to pay fixed (TBAC, 2021). To accommodate this demand, dealers and fixed-income hedge funds were "long swap spreads," receiving the fixed swap rate and taking offsetting short positions in Treasuries (Duarte et al., 2006). Indeed, the 1998 Long Term Capital Management crisis revealed that many intermediaries had substantial long positions in this swap spread trade (Lowenstein, 2000).

This pattern is illustrated in Panel A of Figure 1, which plots 30-year swap spreads alongside primary dealers' net position in U.S. Treasuries from 2001 to 2020. The figure shows that dealers were net short Treasuries during the pre-2008 period when swap spreads were positive. Our interpretation is that this net short Treasury position was driven, to a significant extent, by dealers' long position in the swap spread trade: dealers were receiving the fixed swap rate on net and had a net short position in Treasuries to hedge the resulting interest-rate risk.

In late 2008, long-dated LIBOR swap spreads turned negative and have remained negative ever since, leading to an apparent arbitrage. Spreads' move into negative territory occurred against the backdrop of three major changes in the swap market. First, there was a significant shift in the end-user demand, with net end-user demand swinging from a desire to pay fixed to one to receive fixed (TBAC, 2021). Second, intermediaries became far more concerned with husbanding their capital. The resulting rise in the shadow value of intermediary capital has led to a noteworthy rise in LoOP deviations in a number of intermediated markets. Finally, there was a substantial increase in outstanding Treasury debt during the GFC and, historically, increases in Treasury supply have tended to reduce the money-like convenience premium in Treasury yields and hence swap spreads. We now detail these three forces in turn.

First, as shown in Panel A of Figure 1, primary dealers' position in Treasuries swung from net short to net long in early 2009—just around the time when 30-year swap spreads turned negative. Our interpretation is that net end-user demand swung from a desire to pay fixed to a desire to receive fixed. Thus, dealers' growing net long position in Treasuries was driven by the growing short position they were taking in the swap spread trade. What led to this swing in end-user net demand at the height of the GFC? TBAC (2021) points to a large decline in the demand to pay fixed from hedged mortgage investors. A complementary explanation comes from Klingler and Sundaresan (2019) who argue that a rise in pension underfunding during the GFC led pensions to increasingly receive fixed to manage their duration gaps.

Second, during the GFC, many intermediaries suffered large losses and began husbanding their scarce risk-bearing capital. Motivated by a desire to safeguard financial stability in the aftermath of the GFC, regulators have subjected large dealer banks to more stringent capital regulations. Because it is expensive for dealer banks to finance themselves with equity instead of debt (Boyarchenko et al., 2020), these heightened capital requirements have increased the cost of intermediation. As a result, the shadow value of intermediary capital has increasingly been impounded into market prices since the GFC, leading to a rise in persistent deviations from the LoOP in a range of intermediated markets.[6] (Although the rising shadow value of capital shapes the magnitude of LoOP violations, the *sign* of these violations is naturally linked to the sign of intermediaries' positions.) Most importantly, U.S. regulators introduced the Supplementary Leverage Ratio (SLR) in early 2014 after six years of public discussion. The SLR requires large dealer banks to have Tier 1 capital equal to 6% of their "total leverage exposure," defined as the sum of on-balance-sheet assets plus an adjustment for off-balance-sheet exposures. Unlike traditional forms of risk-based capital regulation, the SLR depends only on the notional scale of dealers' exposures and not on their assessed risk. The SLR has become the binding capital constraint for most large dealers (Duffie, 2017; Greenwood et al., 2017; Boyarchenko et al., 2020).

Finally, there was a substantial increase in outstanding Treasury debt following the onset of the GFC. Specifically, the ratio of marketable Treasury debt to GDP rose from 31% at the end of 2007 to 59% at the end of 2010. Historically, increases in Treasury supply have tended to reduce swap spreads, arguably because this reduces the money-like safety or liquidity premium commanded by Treasuries (Cortes, 2003; Krishnamurthy and Vissing-Jorgensen, 2012). As a result, the large post-GFC expansion in Treasury supply may have sated this special demand for Treasuries, largely eliminating the convenience premium on Treasuries and leading to a decline in swap spreads.

3 Theory

In our model, end users demand long-term interest rate swaps that are supplied by risk-averse and leverage-constrained intermediaries who specialize in the swap market. Fluctuations in swap spreads reflect the interplay between end-user demand and intermediary supply. Because intermediaries face binding leverage constraints, this remains true even if the short-rate differential is always zero and non-zero swap spreads are a failure of the LoOP.

3.1 Setting

Time is discrete, infinite, and is indexed by t. We begin by considering a perpetual interest rate swap and a perpetual Treasury bond that have the same duration. After exploring the determinants of the resulting perpetual swap spread, we extend the model to consider the entire term structure of swap spreads. To maintain tractability, we substitute log returns for ordinary returns throughout and use the Campbell and Shiller (1988) linearization of log returns. These linearity-generating modelling devices do not affect our economic conclusions.

The swap spread trade. Let y_t^S denote the log fixed rate on this perpetual swap at time t and i_t^S the log short-term rate referenced by this swap—e.g., the London Interbank Offer Rate for a LIBOR swap or the Secured Overnight Financing Rate for a SOFR swap. The log excess return

[6] See Gârleanu and Pedersen (2011) for a theoretical analysis of how binding intermediary leverage constraints can lead to LoOP violations; Du et al. (2018) for a detailed study of one prominent post-GFC LoOP violation, namely the deviations from covered interest parity observed in the foreign exchange market; and Siriwardane et al. (2021) for an investigation of the commonalities in various LoOP violations.

on a receive-fixed interest rate swap from t to $t+1$ is approximately

$$r_{t+1}^{S} \equiv \left(y_{t}^{S}-i_{t}^{S}\right)-\frac{\delta}{1-\delta}\left(y_{t+1}^{S}-y_{t}^{S}\right), \tag{1}$$

where $\delta \in (0,1)$ and $1/(1-\delta)$ represents the duration of the swap.[7] The excess return on this receive-fixed swap consists of a carry term $y_t^S - i_t^S$ and a capital gain term $-(\delta/(1-\delta))\left(y_{t+1}^S - y_t^S\right)$ that arises from any change in the swap rate from t to $t+1$. Similarly, the log excess return on a position in perpetual Treasury bonds of the same duration with log yield y_t^T that is financed at the secured short-term financing rate applicable to Treasuries i_t^T —i.e., the rate on repurchase agreements backed by Treasury bonds—is approximately

$$r_{t+1}^{T} \equiv \left(y_{t}^{T}-i_{t}^{T}\right)-\frac{\delta}{1-\delta}\left(y_{t+1}^{T}-y_{t}^{T}\right). \tag{2}$$

The perpetual swap spread is defined as the difference between the fixed rate on perpetual interest rate swaps and the yield on perpetual Treasury bonds: $s_t \equiv y_t^S - y_t^T$. The log excess return from t to $t+1$ on the *receive-fixed swap spread trade* that receives the fixed swap rate and hedges the associated interest-rate risk by going short Treasury bonds is

$$r_{t+1}^{s} \equiv r_{t+1}^{S}-r_{t+1}^{T}=\left(s_{t}-m_{t}\right)-\frac{\delta}{1-\delta}\left(s_{t+1}-s_{t}\right), \tag{3}$$

where $m_t \equiv i_t^S - i_t^T$ is the *short-rate differential* —i.e., the spread between the short-term rate referenced by the swap i_t^S and the short-term Treasury financing rate i_t^T.[8] Our focus is on the relative valuation of swaps and Treasuries—i.e., on the equilibrium level of swap spreads, s_t, and the expected returns on the receive-fixed swap spread trade, $\mathrm{E}_t[r_{t+1}^s]$. Thus, we think of i_t^S, i_t^T, and the general level of long-term rates as summarized by y_t^T as being exogenously given and pinned down by forces outside of our model.[9]

The short-rate differential, $m_t \equiv i_t^S - i_t^T$, governs the fundamental component (if any) of swap spreads. We assume that $m_t = \overline{m} + z_t^m$, where $\overline{m} \equiv \mathrm{E}[m_t] \geq 0$ is the unconditional mean of the short rate differential and z_t^m is a mean-zero state variable that captures time-variation in the short rate differential. A non-zero m_t might either derive from the fact that (i) the short-term interest rate referenced by the swap i_t^S contains compensation for credit risk—as with 3-month LIBOR—or that (ii) short-term Treasury rates embed a special money-like convenience premium relative to other money-market rates. In both cases, m_t would fluctuate over time and we would have $m_t \equiv i_t^S - i_t^T \geq 0$ almost surely, as indeed has been the case historically.

While we allow the cashflow fundamental $m_t \equiv i_t^S - i_t^T$ to differ from zero and to fluctuate stochastically, we will often emphasize the case where $m_t = 0$ for all t almost surely, as would

[7]Assuming this perpetuity makes a series of geometrically-declining fixed payments, δ is governed by the rate at which the payments decline over time. This approximation appears in Campbell (2018) and is an approximate generalization of the fact that the log-return on a n-period zero-coupon bondis *exactly* $r_{t+1}^{(n)} = ny_t^{(n)} - (n-1)\, y_{t+1}^{(n-1)}$.

[8]Naturally, the return on the *pay-fixed swap spread trade* that pays the fixed swap rate and hedges the associated interest-rate risk by going long Treasury bonds is $-r_{t+1}^s$.

[9]It would be easy to endogenize both swaps spreads and the general level of long-term rates merging our model of swap spreads with a Vayanos and Vila (2021)-style model of the term structure of interest rates. In that case, one would need to clear the markets for both long-term Treasuries and for swaps.

(approximately) be the case for a SOFR swap. In this limiting $m_t = 0$ case and in the absence of balance-sheet frictions, the equilibrium swap spread would be zero by the LoOP.

Swap market intermediaries. At time t, risk-averse and leverage-constrained intermediaries who specialize in the swap market allocate their scarce capital w_t between the swap spread arbitrage trade and a risky outside investment opportunity. In practice, we associate these intermediaries with swap desks at broker-dealers or swap traders at fixed-income hedge funds. Since our intermediaries specialize in the swap market, it is natural to think of this outside investment opportunity as another relative-value fixed-income trade. The excess return on this outside investment opportunity is r^o_{t+1}, and we assume that its first two moments are exogenously given by $\mathrm{E}_t\left[r^o_{t+1}\right] = \overline{r}_o > 0$ and $\mathrm{Var}_t\left[r^o_{t+1}\right] = \sigma^2_o > 0$.[10]

For simplicity, we assume an overlapping generations structure where date-t intermediaries are born with capital equal to $w_t = \overline{w} + z^w_t$, where $\overline{w} \equiv \mathrm{E}\left[w_t\right] > 0$ is the unconditional mean of intermediary capital and z^w_t is an exogenous mean-zero state-variable that captures time-variation in intermediary capital.[11] Date-t intermediaries have mean-variance preferences over their one-period ahead wealth $w_{t,\,t+1}$, and their coefficient of absolute risk-aversion is $\alpha \geq 0$.

More formally, letting x_t denote intermediaries' position in the receive-fixed swap spread trade and o_t their position in the outside investment opportunity, date-t intermediaries solve:

$$\max_{x_t,\,o_t}\left\{\mathrm{E}_t\left[w_{t,\,t+1}\right] - \frac{\alpha}{2}\mathrm{Var}_t\left[w_{t,\,t+1}\right]\right\}, \tag{4}$$

subject to the budget constraint

$$w_{t,\,t+1} = w_t + x_t r^s_{t+1} + o_t r^o_{t+1} \tag{5}$$

and the leverage constraint

$$\kappa_x \left|x_t\right| + \kappa_o \left|o_t\right| \leq w_t. \tag{6}$$

Here $\kappa_x, \kappa_o \geq 0$ are the capital requirements associated with the swap spread trade and the outside investment opportunity, respectively. In order to undertake a swap spread trade of notional size $|x_t|$ intermediaries must commit $\kappa_x\left|x_t\right|$ of their scarce capital. For example, according to Boyarchenko et al. (2020), a large U.S. dealer bank who is subject to the Supplementary Leverage Ratio requirement would have $\kappa_x \approx 6.4\%$ for a swap spread trade. We assume the required amount of capital is (approximately) the same whether the intermediary has a long or short position in these trades,

[10] In Internet Appendix B.2 we provide an alternative specification for the outside investment opportunity.

[11] This means that we are not modelling the mechanism through which initial losses on the swap spread trade could be amplified because these losses reduce intermediaries' net worth and tighten future leverage constraints as in Brunnermeier and Pedersen (2009). We could model this amplification mechanism if we assumed that intermediary wealth evolved endogenously according to $w_{t+1} = w_t + x_t r^s_{t+1} + o_t r^o_{t+1}$. This extension would add significant complexity—s_t would depend on w_t as in our model with exogenous wealth shocks, but w_t would also depend endogenously on s_t through the law of motion for intermediary wealth—and an additional source of non-linearity beyond those considered below. Instead, in line with a "segmented-market" view of intermediation, we think of the specialized intermediary as a swap desk within a larger financial institution. In each period the swap desk is endowed with trading capital that depends on the total capital of the financial institution. However, the profits and losses on the swap spread trade have only a small effect on this total capital. Thus, we interpret z^w_t as shifts in the overall balance sheet strength of the intermediary sector that are largely exogenous from the perspective of the specialized swap desk.

hence the absolute values in (6).

End-user demand for swaps. End users of interest rate swaps are agents who demand exposure to long-term interest rates, but who for regulatory, accounting, or other frictional reasons prefer to obtain their desired exposure using swaps as opposed to Treasuries. Intermediaries only need to accommodate the net demand from end users to receive the fixed swap rate—i.e., the demand to receive fixed that is not offset by other end-user demand to pay fixed. Since end users can substitute between swaps and Treasuries, we allow net demand to receive fixed to be increasing in swap spreads. Specifically, we assume net end-user demand to receive the fixed swap rate is

$$d_t = \overline{d} + z_t^d + \gamma s_t, \tag{7}$$

where $\gamma \geq 0$ and z_t^d is a mean-zero state variable that captures shifts in end-user demand.

State variable dynamics. We assume an autoregressive process for the three state variables governing the short rate differential (z_t^m), end-user demand (z_t^d), and intermediary wealth (z_t^w). Specifically, we assume the vector of state variables $\mathbf{z}_t = [z_t^m, z_t^d, z_t^w]'$ follows

$$\mathbf{z}_{t+1} = \boldsymbol{\varrho}\mathbf{z}_t + \boldsymbol{\varepsilon}_{t+1}, \tag{8}$$

where $\boldsymbol{\varrho} = diag\left(\rho_m, \rho_d, \rho_w\right)$ is a diagonal matrix of the AR(1) coefficients $\rho_m, \rho_d, \rho_w \in [0, 1)$ and $\boldsymbol{\varepsilon}_{t+1} = [\varepsilon_{t+1}^m, \varepsilon_{t+1}^d, \varepsilon_{t+1}^w]'$ is the vector of structural shocks. For simplicity, we assume the three structural shocks are orthogonal to each other: $\text{Var}_t[\boldsymbol{\varepsilon}_{t+1}] = diag\left(\sigma_m^2, \sigma_d^2, \sigma_w^2\right)$, where $\sigma_i^2 \equiv \text{Var}\left[\varepsilon_{t+1}^i\right]$ for $i = m, w, d$.

Market clearing. Since interest rate swaps are in zero net supply, market clearing requires

$$x_t + d_t = 0. \tag{9}$$

In particular, if the net demand from end users to receive the fixed swap rate is positive, $d_t > 0$, then in equilibrium intermediaries must take on a short position in the receive-fixed swap spread trade (paying the fixed swap rate and going long Treasuries) equal to $x_t = -d_t < 0$. By contrast, if there is net end-user demand to pay the fixed swap rate, intermediaries must take on a long position in the receive-fixed swap spread trade (receiving the fixed swap rate and going short Treasuries).

3.2 Equilibrium swap spreads

We begin by providing a general characterization of swap spreads that is applicable in settings where end-user net demand may change signs over time and where intermediary leverage constraints may only bind periodically. In this general case, swap spreads will be a nonlinear function of intermediary wealth w_t and end-user net demand d_t. This general characterization sheds light on the long-run history of the swap market and the regime change it experienced during the 2008 Global Financial Crisis. We then provide an affine characterization of swap spreads that is valid in settings where (i) the sign of end-user demand is constant over time and (ii) the leverage constraint is always binding. The resulting affine model is particularly helpful in understanding the post-Global Financial Crisis (GFC) period that is the focus of our empirical analysis. Finally, we discuss how the predictions from the affine model generalize once we introduce nonlinearities by relaxing assumptions (i) and (ii).

3.2.1 General characterization

Letting $\psi_t \geq 0$ denote the Lagrange multiplier associated with the leverage constraint (6) at time t and assuming $\text{Cov}_t[r^s_{t+1}, r^o_{t+1}] = 0$ for simplicity, intermediaries' first-order condition for x_t is

$$\text{E}_t[r^s_{t+1}] = \alpha V_t \cdot x_t + \kappa_x sgn\left(x_t\right) \cdot \psi_t, \tag{10}$$

where $V_t \equiv \text{Var}_t[r^s_{t+1}] = \left(\frac{\delta}{1-\delta}\right)^2 \text{Var}_t[s_{t+1}]$ is the conditional variance of r^s_{t+1}.[12] Similarly, the first order condition for o_t is

$$\overline{r}_o = \alpha\sigma_o^2 \cdot o_t + \kappa_o sgn\left(o_t\right) \cdot \psi_t. \tag{11}$$

Since $\overline{r}_o > 0$ and $\psi_t \geq 0$, we must have $o_t \geq 0$, implying that the Lagrange multiplier is

$$\psi_t = \frac{\alpha\sigma_o^2}{\kappa_o}\left(\frac{\overline{r}_o}{\alpha\sigma_o^2} - o_t\right) \geq 0. \tag{12}$$

Thus, the shadow value of intermediary capital is proportional to the difference between the unconstrained investment in the outside opportunity, $\overline{r}_o/\left(\alpha\sigma_o^2\right)$, and intermediaries' current investment, o_t. Combining the leverage constraint in (6) and (12), we have

$$\psi_t = \psi\left(\left|x_t\right|, w_t\right) = \max\left\{0, \frac{\alpha\sigma_o^2}{\kappa_o^2}\left(\kappa_o\frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x\left|x_t\right| - w_t\right)\right\}. \tag{13}$$

Naturally, the shadow value of intermediary capital is greater when intermediary capital w_t is lower and when the scale of intermediaries' position in the swap spread trade $|x_t|$ is larger.

Combining (3), (10), and (13), and imposing market clearing (9), we find that the equilibrium expected return on the receive-fixed swap spread trade, $\text{E}_t[r^s_{t+1}]$, satisfies:

$$\overbrace{\left(s_t - m_t\right) - \frac{\delta}{1-\delta}\left(\text{E}_t\left[s_{t+1}\right] - s_t\right)}^{\text{E}_t[r^s_{t+1}]} = \overbrace{\left(-\kappa_x\right)sgn\left(d_t\right) \cdot \psi\left(\left|d_t\right|, w_t\right)}^{\text{Compensation for using scarce capital}} + \overbrace{\left(-\alpha\right)V_t \cdot d_t}^{\text{Compensation for risk}}. \tag{14}$$

Equation (14) highlights the two key forces that shape the equilibrium expected returns on the swap spread trade: compensation for using scarce intermediary capital and compensation for risk. When $\kappa_x > 0$, constrained intermediaries will require compensation for committing their scarce capital to the swap spread trade even if this trade is completely riskless. And, being risk-averse, specialized intermediaries will require additional compensation for bearing the risk that they may suffer losses due to unexpected changes in swap spreads.

[12] We assume $\text{Cov}_t[r^s_{t+1}, r^o_{t+1}] = 0$ for simplicity. Relaxing this assumption adds hedging terms proportional to $\text{Cov}_t[r^s_{t+1}, r^o_{t+1}]$, which complicate the resulting expressions without qualitatively changing our conclusions.

Since in equilibrium intermediaries must take positions that are equal in size and opposite in sign to those of end users, the equilibrium expected return on a receive-fixed swap spread position, $\mathrm{E}_t[r^s_{t+1}]$, has the opposite sign of the net end-user demand to receive the fixed rate, d_t. For instance, if net demand to receive fixed is negative—as was arguably the case prior to the GFC—we must have $\mathrm{E}_t[r^s_{t+1}] > 0$ to induce intermediaries to take the required long position in the receive-fixed swap spread trade. By contrast, if net demand to receive fixed is positive—as seems to have been the case since the GFC—we must have $\mathrm{E}_t[r^s_{t+1}] < 0$ to induce intermediaries to take the required short position in the receive-fixed spread trade. Thus, while a rise in intermediary balance-sheet costs $\psi\left(|d_t|, w_t\right)$ will raise the absolute magnitude of $\mathrm{E}_t[r^s_{t+1}]$, equation (14) clarifies that a rise in balance-sheet costs by itself will not lead $\mathrm{E}_t[r^s_{t+1}]$ to switch sign.

Iterating (14) forward and assuming $\lim_{k\to\infty}\delta^k \mathrm{E}_t\left[s_{t+k}\right] = 0$—as would be the case in any stationary environment—we find that the equilibrium level of swap spreads is

$$\begin{aligned} s_t = &\overbrace{(1-\delta)\sum\nolimits_{k=0}^{\infty}\delta^k \mathrm{E}_t\left[m_{t+k}\right]}^{\text{Expected short-rate differentials}} + \overbrace{(1-\delta)\sum\nolimits_{k=0}^{\infty}\delta^k \mathrm{E}_t\left[(-\kappa_x)\, sgn\left(d_{t+k}\right)\psi\left(|d_{t+k}|, w_{t+k}\right)\right]}^{\text{Expected compensation for using scarce capital}} \\ &+ \underbrace{(1-\delta)\sum\nolimits_{k=0}^{\infty}\delta^k \mathrm{E}_t\left[(-\alpha)\, V_{t+k} d_{t+k}\right]}_{\text{Expected compensation for risk}}. \end{aligned} \tag{15}$$

The first term on the right-hand side of (15) is the fundamental component of swap spreads. Recalling that $m_t \equiv i^S_t - i^T_t$, this is simply the expected future difference between the short-term rate referenced by the swap (i^S_t) and the secured Treasury financing rate (i^T_t) averaged over the lifetime of the swap. Under the assumption that $m_t > 0$ almost surely as would be the case for LIBOR-based swaps, this fundamental term pushes towards having positive swap spreads. However, this term is (approximately) zero for SOFR swaps where $m_t \equiv 0$.

The second term is the expected future compensation for consuming scarce intermediary capital over the life of the swap. Since $\psi_t = \psi\left(w_t, |d_t|\right) \geq 0$, this term has the opposite sign of net end-user demand (d_t). For instance, if the net demand to receive the fixed swap rate is positive, this term pushes towards having negative swap spreads.

The third and final term is the expected future compensation for bearing the risk associated with swap spread volatility over the life of the swap. Like the second term, this final term has the opposite sign of net end-user demand to receive fixed.

3.2.2 Affine equilibrium

To derive an affine equilibrium that we can readily take to the post-GFC data, we make two substantive economic assumptions:

A1: *End-user net demand to receive the fixed swap rate is always positive*. Formally, we assume $0 < d_t$ at all dates.

A2: *Intermediaries' leverage constraint in equation (6) always binds*. Formally, we assume

$$w_t < \kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x d_t$$

at all dates, ensuring that we always have $\psi_t > 0$.

Assumption A1 rules out the nonlinearities that arise if end-user net demand can switch signs over time. This assumption is consistent with prime dealers' net positioning in long-term Treasuries and the negative swap spreads that have been observed since 2008; see Panel A of Figure 1. By contrast, assumption A2 rules out the nonlinearities that arise if leverage constraints only bind occasionally. Below we will consider nonlinearities that arise if we relax A1 and/or A2.

When end-user demand is perfectly inelastic ($\gamma = 0$), we require an additional technical assumption to ensure that market clearing is always feasible:

A3: *Intermediaries always have sufficient capital to accommodate end-user demand in the swap market*. Formally, we assume $\kappa_x |d_t| < w_t$ at all dates.

However, assumption A3 is not required when $\gamma > 0$. To ensure that A1, A2, and A3 (if necessary) hold, we assume that the shocks $\boldsymbol{\varepsilon}_t$ have a bounded support. Given the assumed AR(1) dynamics, this implies that the state variables $\mathbf{z}_t$ also have a bounded support. And, when $(\kappa_o \overline{r}_o) / (\alpha\sigma_o^2) > 0$, we can choose $\overline{m}$, $\overline{d}$, and $\overline{w}$ so that A1, A2, and A3 (if necessary) hold at each point in the state space. See Internet Appendix B.3 for additional details.

We conjecture that equilibrium swap spreads s_t are an affine function of the state vector $\mathbf{z}_t$:

$$s_t = A_0 + A_m z_t^m + A_d z_t^d + A_w z_t^w. \tag{16}$$

Under assumptions A1, A2, A3 (if necessary), and assuming that spreads take the conjectured affine form in (16), $V = \mathrm{Var}_t[r_{t+1}^s]$ is constant over time and condition (14) becomes

$$\overbrace{(s_t - m_t) - \frac{\delta}{1-\delta}(\mathrm{E}_t[s_{t+1}] - s_t)}^{\mathrm{E}_t[r_{t+1}^s]} = (-\kappa_x) \cdot \overbrace{\frac{\alpha\sigma_o^2}{\kappa_o^2}\left(\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x d_t - w_t\right)}^{\psi_t} + (-\alpha)V \cdot d_t. \tag{17}$$

A rational expectations equilibrium of our model is a fixed point of an operator which gives the price-impact coefficients A_m, A_d, and A_w that clear the swap market when intermediaries believe the risk of the swap spread arbitrage trade is determined by some initial set of price-impact coefficients. And, this can be recast as a scalar fixed-point problem involving $V = \mathrm{Var}_t[r_{t+1}^s]$. Combining the conjectured affine form (16), the end-user demand curve (7), and the equilibrium condition (17), we obtain the following result:

Theorem 1 *In the affine equilibrium that may exist when assumptions A1, A2, and A3 (if neces-*

sary) hold, we have

$$A_0 = \frac{\overline{m} - \kappa_x \frac{\alpha\sigma_o^2}{\kappa_o^2}\left[\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x \overline{d} - \overline{w}\right] - \alpha V \overline{d}}{1 + \alpha\gamma\left[\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 + V\right]}, \tag{18a}$$

$$A_m = \frac{1}{\frac{1-\rho_m\delta}{1-\delta} + \alpha\gamma\left[\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 + V\right]} > 0, \tag{18b}$$

$$A_d = -\frac{\alpha\left[\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 + V\right]}{\frac{1-\rho_d\delta}{1-\delta} + \alpha\gamma\left[\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 + V\right]} \leq 0, \tag{18c}$$

$$A_w = \frac{\frac{1}{\kappa_x}\alpha\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2}{\frac{1-\rho_w\delta}{1-\delta} + \alpha\gamma\left[\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 + V\right]} \geq 0, \tag{18d}$$

and

$$V = \left(\frac{\delta}{1-\delta}\right)^2 \left(A_m^2\sigma_m^2 + A_d^2\sigma_d^2 + A_w^2\sigma_w^2\right) \geq 0. \tag{19}$$

Equations (18b)-(18d) and (19) together define a higher-order polynomial equation in V. When either $\gamma > 0$ or $\alpha\sigma_d^2 = 0$, a solution to this equation always exists. When $\gamma = 0$ and $\alpha\sigma_d^2 > 0$, a solution only exists if intermediary risk aversion α is below a threshold $\alpha^* > 0$. When $\alpha\sigma_d^2 = 0$, the equilibrium is unique. When $\alpha\sigma_d^2 > 0$, there are generally multiple solutions corresponding to multiple affine equilibria. However, there is at most one equilibrium solution that is (i) stable in the sense that it is robust to small perturbations in intermediaries' beliefs about the price impact of demand shocks and (ii) does not diverge in the limit where demand risk vanishes ($\sigma_d^2 \to 0$).

To provide intuition for Theorem 1, we consider a number of special cases.

With a positive short-term interest rate differential ($m_t > 0$ for all t) and risk-neutral and unconstrained intermediaries ($\alpha = \kappa_x = 0$), the model has a unique equilibrium with positive swap spreads. In this case, the swap spread is simply the expected short-rate differential averaged over the life of the swap. Moreover, if short-rate differentials vary over time ($\sigma_m^2 > 0$) and risk-averse but unconstrained intermediaries must absorb end-user demand shocks ($\alpha > 0$, $\kappa_x = 0$, and $\sigma_d^2 > 0$), these demand shocks will shift the required compensation for bearing swap spread risk. This means that spreads will also reflect the expected risk compensation over the life of the swap. In either case, short rate differentials are akin to cash flow fundamentals: they do not induce LoOP violations and have standard effects on swap spreads.

Therefore, to illustrate the more novel forces at work in our model, we assume the short rate differential is always zero ($m_t = 0$ for all t) in the following examples. That is, we consider SOFR swaps where the LoOP offers the stark prediction that swap spreads should always be zero. We also assume intermediaries are risk averse ($\alpha > 0$), there are shocks to intermediary wealth ($\sigma_w^2 > 0$), and end-user demand is completely inelastic ($\gamma = 0$).

When end-user demand is stochastic ($\sigma_d^2 > 0$) but swap spread trades do not consume scarce capital ($\kappa_x = 0$), the model can have two affine equilibria: (i) a "zero-volatility" equilibrium in which SOFR swap spreads are always zero and (ii) a "high-volatility" equilibrium in which SOFR swap spreads are driven by fluctuations in end-user demand. This special case is similar to De Long et al. (1990). When demand is stochastic, the perceived risk of the swap spread trade depends on how intermediaries believe spreads will react to future demand shocks, opening the door to multiple equilibrium solutions.[13] For example, if intermediaries believe that future demand shocks will have an impact on swap spreads, they will perceive the swap spread trade as being risky. As a result, intermediaries will only absorb demand shocks if they are compensated by spread changes and positive expected returns, making their initial belief self-fulfilling. Conversely, if intermediaries believe demand shocks will have no impact on spreads, they will perceive the swap spread trade as being riskless. As a result, intermediaries will elastically absorb demand shocks even though spreads are always zero. However, only the equilibrium in which SOFR swap spreads are always zero is (i) stable in the sense that it is robust to small perturbations in intermediaries' beliefs about the price impact of demand shocks and (ii) does not diverge in the limit where demand risk vanishes ($\sigma_d^2 \to 0$). Thus, there is a clear sense in which the equilibrium where LoOP holds is the natural outcome in this case.

When swap spread trades consume scarce intermediary capital ($\kappa_x > 0$) but there is no demand risk ($\sigma_d^2 = 0$), the model has a unique affine equilibrium in which fluctuations in SOFR swap spreads are driven by time-variation in intermediary capital. Intuitively, the shadow value of capital is positive ($\psi_t > 0$) and, since intermediaries' (constant) swap spread positions consume scarce capital ($\kappa_x > 0$), swap spreads are non-zero and the LoOP fails.

With both demand risk ($\sigma_d^2 > 0$) and balance-sheet frictions ($\kappa_x > 0$), the model can have two affine equilibria—one with low swap spread volatility and one with higher volatility—but only the low-volatility equilibrium is stable. However, this low-volatility equilibrium is no longer trivial. In this equilibrium, intermediaries understand that binding leverage constraints will prevent them from enforcing LoOP, leading to non-zero SOFR swap spreads. As a result, shocks to end-user demand and intermediary wealth will impact future spreads, creating demand-supply imbalance risk for which risk-averse intermediaries must be compensated. Thus, unlike in De Long et al. (1990), LoOP violations and arbitrage risk arise even in our model's stable equilibrium.

Finally, our conclusions hold with a few modifications when end-user demand is elastic ($\gamma > 0$). When demand is elastic, an equilibrium always exists. And, while there can be multiple equilibria when $\sigma_d^2 > 0$, there is a unique equilibrium that is both stable and that does not diverge in the limit where $\sigma_d^2 \to 0$.[14]

In conclusion, our model features multiplicity due to the possibility of self-fulfilling conjectures about the price-impact of demand shocks. In what follows, we assume an equilibrium exists (e.g., it is sufficient to assume $\gamma > 0$) and we focus on the unique equilibrium that is stable and

[13] When $\sigma_m^2 = 0$, $\kappa_x = 0$, and $\sigma_d^2 > 0$, the system of equations reduces to a quadratic equation in V. The stable solution is $V = 0$ and the other solution has $V > 0$. More generally, when $\gamma = 0$ and $\sigma_d^2 > 0$, the system reduces to a quadratic in V, which has real roots so long as α is not too large.

[14] When $\gamma > 0$, the fixed-point problem is equivalent to finding the roots of a 7th-order polynomial in V in the general case where $\alpha > 0$, $\kappa_x > 0$, $\sigma_d^2 > 0$, $\sigma_w^2 > 0$, and $\sigma_m^2 > 0$. This reduces to 5th-order polynomial when $\sigma_m^2 = 0$. We can show that this polynomial always has at least one positive root and it only has one root that is both stable and remains in the limit where $\sigma_d^2 \to 0$.

does not diverge when demand risk vanishes. In this equilibrium, swap spreads are less volatile and demand shocks have a smaller price impact as compared to the other candidate equilibria. Nevertheless, spreads are non-zero, which reflects the combined effects of short rate differentials, binding leverage constraints, and demand-and-supply imbalance risk.

3.2.3 Predictions from the affine model

Since the post-GFC period will be the main focus of our empirical analysis, in this section we highlight the predictions that emerge in the affine equilibrium that obtains when end-users always demand to receive fixed on net and intermediaries' leverage constraint always binds—i.e., when assumptions A1, A2, and A3 (if necessary) hold. Furthermore, we assume that α, κ_x, σ_d^2, $\sigma_w^2 > 0$.

Average level of swap spreads. Our model pinpoints the forces that determine the time-series average level of swap spreads:

Proposition 1 *The time-series average level of long-term swap spreads is given by*

$$A_0 = \overline{m} - \kappa_x \cdot \overbrace{\frac{\alpha\sigma_o^2}{\kappa_o^2}\left[\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x E\left[d_t\right] - \overline{w}\right]}^{E[\psi_t]} - \alpha V \cdot \overbrace{\left(\overline{d} + \gamma A_0\right)}^{E[d_t]},$$

Since $E\left[\psi_t\right] > 0$ and $E\left[d_t\right] > 0$ by assumption A1, long-term swap spreads are negative on average ($A_0 < 0$) if the average short rate differential satisfies $\overline{m} < \overline{m}^*$ for some constant $\overline{m}^* > 0$ that is given in the Appendix.

Just as in (15), the time-series average level of swap spreads (A_0) is determined by three forces: the average short rate differential ($\overline{m}$), the average compensation for committing scarce capital to the swap spread trade ($-\kappa_x \cdot \mathrm{E}\left[\psi_t\right]$), and the average compensation for bearing swap spread risk ($-\alpha V \cdot \mathrm{E}\left[d_t\right]$). The average spread is increasing in the average short rate differential, $\overline{m}$, which is positive and non-negligible for LIBOR swaps, is negligible for OIS swaps, and is (approximately) zero for SOFR swaps. Since assumptions A1 and A2 imply $\mathrm{E}\left[\psi_t\right] > 0$ and $\mathrm{E}\left[d_t\right] > 0$, the other two forces—compensation for committing scarce capital and compensation for risk—push swap spreads into negative territory on average. Thus, our model predicts negative average spreads for long-dated OIS and SOFR swaps in the post-GFC era. Relative to those, the short rate differential will push the average level of long-dated LIBOR swap spreads towards positive territory. However, so long as $\overline{m}$ is not too large, balance sheet and risk considerations will dominate, and long-dated LIBOR swap spreads will also be negative on average.

Expected returns on the swap spread trade. Our model emphasizes that two distinct forces—compensation for consuming scarce intermediary capital and compensation for bearing swap spread risk—drive the expected returns on the swap spread trade. Specifically, (17) implies the following result:

Proposition 2 *The expected return on the received-fixed swap spread trade is $E_t[r^s_{t+1}] = -\kappa_x \cdot \psi_t - \alpha V \cdot d_t$, where $\psi_t = \left(\alpha\sigma_o^2/\kappa_o^2\right)\left(\kappa_o\overline{r}_o/\left(\alpha\sigma_o^2\right) + \kappa_x d_t - w_t\right)$ is the shadow value of intermediary* *capital. Thus, $\mathrm{E}_t[r^s_{t+1}]$ is decreasing in end-user demand d_t (i.e., increasing in intermediaries' position x_t), even controlling for the shadow value of capital ψ_t.*

In equilibrium, the expected returns on the swap spread trade must compensate intermediaries for the opportunity cost of committing their scarce capital to this trade ($-\kappa_x \cdot \psi_t < 0$) as well as for bearing swap spread risk ($-\alpha V \cdot d_t < 0$). As end-user demand d_t to receive fixed increases, the expected return on the receive-fixed spread trade needs to become more negative to induce intermediaries to take a larger short position in this trade. Furthermore, since intermediaries require compensation for bearing swap spread risk, this is true even after controlling for fluctuations in the compensation that they require for using their scarce capital.

The model also characterizes the effects that shifts in the three underlying structural factors—the short rate differential (z^m_t), end-user demand (z^d_t), and intermediary capital (z^w_t)—have on expected returns $\mathrm{E}_t[r^s_{t+1}]$. Using Theorem 1, we obtain the following results:

Proposition 3 *The expected return on the receive-fixed swap spread trade is*

$$\mathrm{E}_t[r^s_{t+1}] = B_0 + B_m z^m_t + B_d z^d_t + B_w z^w_t, \tag{20}$$

where $B_0 = A_0 - \overline{m} < 0$, $B_m = \frac{1-\delta\rho_m}{1-\delta} A_m - 1 \leq 0$, $B_d = \frac{1-\rho_d\delta}{1-\delta} A_d < 0$, and $B_w = \frac{1-\rho_w\delta}{1-\delta} A_w > 0$. Furthermore, $B_m = 0$ when $\gamma = 0$ but $B_m < 0$ when $\gamma > 0$.

To see the intuition, it is easiest to flip the sign and think about the expected returns on the pay-fixed swap spread position ($-\mathrm{E}_t[r^s_{t+1}]$) that intermediaries hold in equilibrium under assumption A1. To begin, Proposition 3 says that the expected returns on this pay-fixed spread trade are positive on average ($-B_0 = -\mathrm{E}[r^s_{t+1}] > 0$). Next, outward shifts in end-user demand to receive fixed raise the expected returns on the pay-fixed spread trade because they raise both the shadow value of intermediary capital and the required compensation for bearing swap spread risk:

$$-B_d = -\frac{\partial \mathrm{E}_t[r^s_{t+1}]}{\partial z^d_t} = \kappa_x \cdot \overbrace{\frac{\alpha\sigma^2_o}{\kappa^2_o}\kappa_x(1+\gamma A_d)}^{\partial\psi_t/\partial z^d_t > 0} + \alpha V \cdot \overbrace{(1+\gamma A_d)}^{\partial d_t/\partial z^d_t > 0} > 0.$$

By contrast, shifts in intermediary wealth move the shadow value of intermediary capital and the compensation for risk in opposite directions. However, the former effect always dominates, implying that wealth shocks decrease the expected returns on the pay-fixed spread trade:

$$-B_w = -\frac{\partial \mathrm{E}_t[r^s_{t+1}]}{\partial z^w_t} = \kappa_x \cdot \overbrace{\frac{\alpha\sigma^2_o}{\kappa^2_o}(\kappa_x\gamma A_w - 1)}^{\partial\psi_t/\partial z^w_t < 0} + \alpha V \cdot \overbrace{\gamma A_w}^{\partial d_t/\partial z^w_t > 0} < 0.$$

Finally, when end-user demand is inelastic ($\gamma = 0$), changes in the short-rate differential m_t do not alter $\mathrm{E}_t[r^s_{t+1}]$ even though they impact swap cash flows ($B_m = 0$). However, when end-user demand is elastic ($\gamma > 0$), increases in m_t raise swap spreads and hence end-user demand to receive fixed. As a result, an increase in m_t must be accompanied by higher expected return on the pay-fixed spread trade to induce intermediaries to accommodate this demand ($-B_m > 0$).

Identification of demand and supply shocks. Our affine model provides sign restrictions that can help us identify the structural demand (ε_t^d) and supply (ε_t^w) shocks using data on the short rate differential (m_t), the level of swap spreads (s_t), and end-users' net position in swaps (d_t). In particular, we obtain the following result:

Proposition 4 *The short rate differential m_t, the long-term swap spread s_t, and end users' net position in receive-fixed swaps d_t can be written as*

$$\overbrace{\begin{bmatrix} m_t \\ s_t \\ d_t \end{bmatrix}}^{\mathbf{y}_t} = \overbrace{\begin{bmatrix} \overline{m} \\ A_0 \\ \left(\overline{d} + \gamma A_0\right) \end{bmatrix}}^{\mathbf{a}} + \overbrace{\begin{bmatrix} 1 & 0 & 0 \\ A_m & A_d & A_w \\ \gamma A_m & (1+\gamma A_d) & \gamma A_w \end{bmatrix}}^{\mathbf{A}} \overbrace{\begin{bmatrix} z_t^m \\ z_t^d \\ z_t^w \end{bmatrix}}^{\mathbf{z}_t}. \tag{21}$$

Combined with (8), the structural VAR representation of the model's equilibrium is

$$\mathbf{y}_{t+1} = \left(\mathbf{I} - \mathbf{A}\boldsymbol{\varrho}\mathbf{A}^{-1}\right)\mathbf{a} + \mathbf{A}\boldsymbol{\varrho}\mathbf{A}^{-1}\mathbf{y}_t + \mathbf{A}\boldsymbol{\varepsilon}_{t+1}, \tag{22}$$

where $\mathbf{I}$ is the 3×3 identity matrix and $\boldsymbol{\varepsilon}_{t+1} = [\varepsilon_{t+1}^m, \varepsilon_{t+1}^d, \varepsilon_{t+1}^w]'$. We have $A_m > 0$, $A_d < 0$, $A_w > 0$, and $(1 + \gamma A_d) > 0$. Moreover, assuming $\gamma > 0$, we also have $\gamma A_m > 0$ and $\gamma A_w > 0$.

Proposition 4 implies that the matrix $\mathbf{A}$ of structural VAR coefficients can be identified using a combination of sign and zero restrictions. To see the intuition most simply, suppose that $m_t = 0$ always. Assumptions A1 and A2 then imply that we always have $s_t < 0$ and we can think of $|s_t| > 0$ as the price that end users pay intermediaries to supply receive-fixed swaps. Rewriting (7) and (17), end-users' inverse demand function for receive-fixed swaps is

$$|s_t|^{Demand} = \gamma^{-1}\left(\overline{d} + z_t^d - d_t\right), \tag{23}$$

and intermediaries' inverse supply function for receive-fixed swaps is

$$|s_t|^{Supply} = (1-\delta)\left[\overbrace{\underbrace{\kappa_x \cdot \frac{\alpha\sigma_o^2}{\kappa_o^2}\left(\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x d_t - \overline{w} - z_t^w\right)}_{\psi_t} + \alpha V \cdot d_t}^{\left|\mathrm{E}_t[r_{t+1}^s]\right|}\right] + \delta \mathrm{E}_t\left[|s_{t+1}|\right]. \tag{24}$$

Thus, end-user demand is downward-sloping ($\partial\,|s_t|^{Demand}/\partial d_t < 0$) and, holding $\mathrm{E}_t\left[|s_{t+1}|\right]$ fixed, intermediaries' supply is upward-sloping ($\partial\,|s_t|^{Supply}/\partial d_t > 0$). Equilibrium occurs where these two curves intersect in the $(d_t, |s_t|)$ space.

An increase in z_t^d shifts end-users' demand curve outward. Since intermediaries' supply curve is upward sloping, this increase in z_t^d is associated with a rise in both d_t and $|s_t|$. Similarly, an increase in z_t^w shifts intermediaries' supply curve outward. Since end-user demand is downward sloping, this increase in z_t^w is associated with a rise in d_t and a decline in $|s_t|$. In our empirical work, we use this textbook demand-and-supply logic to set-identify the structural demand (ε_t^d)

and supply (ε_t^w) shocks. Specifically, if there are positive (negative) innovations to both d_t and $|s_t|$, then there must have been a positive (negative) shock to end-user demand (ε_t^d). Conversely, if there is positive (negative) innovation to d_t and a negative (positive) innovation to $|s_t|$, there must have been a positive (negative) shock to intermediary supply (ε_t^w).

3.2.4 Allowing for nonlinearities

In Internet Appendix B.3, we provide a more general characterization of swap spreads based on equations (14) and (15) that allows for nonlinearities. For simplicity, we focus on the case where swaps have zero fundamental value ($m_t \equiv 0$) and where end-user demand for swaps is inelastic ($\gamma = 0$), so there are just two exogenous state variables: end-user demand (d_t) to receive the fixed swap rate and intermediary wealth (w_t). We let $s_t = s(d_t, w_t)$ denote the equilibrium level of swap spreads, $\text{E}_t[r^s_{t+1}] = s_t - (\delta/(1-\delta))(\text{E}_t[s_{t+1}] - s_t)$ the expected return on the receive-fixed swap spread trade, and $V_t \equiv \text{Var}_t[r^s_{t+1}] = V(d_t, w_t)$ the corresponding conditional variance.

Two features of our general model can give rise to nonlinearities. First, intermediaries' balance sheet constraint is itself nonlinear, depending on $|d_t|$ instead of d_t: this is relevant if end-user demand—and hence the balance-sheet cost of intermediating swaps—can switch signs. Second, intermediaries' balance sheet constraint may not always bind. This means that the *risk* of the swap spread trade is time-varying. The affine equilibrium that we emphasized in the prior section arises when we make simplifying assumptions that switch off both of these features, namely, that (A1) end-user demand is always positive and (A2) intermediaries' balance-sheet constraint is always binding.

In Internet Appendix B.3, we characterize a number of important special cases of this (potentially) nonlinear model:

- *Case 1*: When both A1 and A2 hold, we obtain Theorem 1 above. Specifically, s_t and $\text{E}_t[r^s_{t+1}]$ are affine functions of d_t and w_t and $V_t = V$ is constant. Furthermore, we have $\partial s_t/\partial d_t = A_d < 0$ and $\partial \text{E}_t[r^s_{t+1}]/\partial d_t \propto A_d < 0$, whereas $\partial s_t/\partial w_t = A_w > 0$ and $\partial \text{E}_t[r^s_{t+1}]/\partial w_t \propto A_w > 0$.

- *Case 2*: When we relax A2 but maintain A1—i.e., if intermediaries' balance-sheet constraint can be slack, but end-user demand is always positive—then s_t, $\text{E}_t\left[r^s_{t+1}\right]$, and V_t are non-linear functions that satisfy $\partial s_t/\partial d_t < 0$ and $\partial s_t/\partial w_t > 0$; $\partial \text{E}_t\left[r^s_{t+1}\right]/\partial d_t < 0$ and $\partial \text{E}_t\left[r^s_{t+1}\right]/\partial w_t > 0$; and, finally, $\partial V_t/\partial d_t > 0$ and $\partial V_t/\partial w_t < 0$.

  Intuitively, even if leverage constraints are not binding at time t, the mere potential for them to bind in the future makes swap spread trades risky for intermediaries, and they will only accommodate end-user demand for long-term swaps if they are compensated for this risk (i.e., we have $V_t > 0$ even when $\psi_t = 0$). Furthermore, even if the constraint is currently slack, fluctuations in end-user demand and intermediary capital will shape the likelihood that the constraint will bind in the future and hence the current level of spreads.

  Thus, the only qualitative difference between *Case 2* and the affine model of *Case 1* is that the risk of the swap spread trade now fluctuates endogenously as a function of d_t and w_t.

Specifically, swap spreads become more volatile when it is more likely that intermediary constraints will bind in the future, explaining why the variance is increasing in end-user demand (d_t) and decreasing in intermediary wealth (w_t). This endogenous-risk effect adds a further channel through which increases in current end-user demand (or reductions in intermediary wealth) raise the expected returns on the swap spread trade and, hence, the current level of swap spreads.

- *Case 3*: The solution can also be characterized rigorously if we relax both A1 and A2, but make the model symmetric by assuming that the steady-state level of end-user demand is zero ($\overline{d} = 0$). When $\overline{d} = 0$, the sign of demand may switch over time, but the expected sign of future demand is always the same as the sign of current demand, making the model quite tractable. In this case, s_t, $\mathrm{E}_t\left[r^s_{t+1}\right]$, and V_t are non-linear functions that satisfy $\partial s_t/\partial d_t < 0$ and $sgn\left(\partial s_t/\partial w_t\right) = sgn\left(d_t\right)$; $\partial \mathrm{E}_t\left[r^s_{t+1}\right]/\partial d_t < 0$ and $sgn\left(\partial \mathrm{E}_t\left[r^s_{t+1}\right]/\partial w_t\right) = sgn\left(d_t\right)$; and, finally, $sgn\left(\partial V_t/\partial d_t\right) = sgn\left(d_t\right)$ and $\partial V_t/\partial w_t < 0$.

  Relative to *Case 2*, there is an additional source of non-linearity due to the fact that end-user demand can switch signs. This means that $sgn\left(s_t\right) = sgn\left(\mathrm{E}_t\left[r^s_{t+1}\right]\right) = -sgn\left(d_t\right)$, and explains why the impact of intermediary wealth on expected returns and spreads depends on the sign of end-user demand. Furthermore, as in *Case 2* above, swap spreads become more volatile when it is more likely that constraints will bind in the future. Since the constraint depends on $|d_t|$ and we assume $\overline{d} = 0$, this explains why variance is decreasing in end-user demand when $d_t < 0$ and increasing in end-user demand when $d_t > 0$.

Internet Appendix B.3 includes several more general results about this (potentially) nonlinear model. Overall, the key qualitative predictions of our affine model, including Propositions 1, 2, and 3 (as well as Proposition 7 in Internet Appendix B.1) generalize naturally once we allow for nonlinearities. While a non-linear version of our identification argument in Proposition 4 holds in *Case 2*, separately identifying the impact of demand and supply becomes more complicated when end-user demand can switch signs. For instance, in *Case 3*, a non-linear version of the identification argument in Proposition 4 only hold once one conditions on $sgn\left(d_t\right)$.[15]

3.3 The term structure of swap spreads

Thus far, we have modelled the spread on a single perpetual long-term swap. To derive predictions for the full term structure of swap spreads, in Internet Appendix B.4 we extend our model by introducing a series of n-period zero-coupon swaps alongside this perpetual swap.[16] As above,

[15]Nonlinear versions of our model, which relax A1 or A2, may also prove useful in thinking about the term structure of Law of One Price deviations in other settings—e.g., the bond-CDS basis or deviations from covered interest rate parity (CIP). For the bond-CDS basis, we suspect that the sign of intermediaries' positions have been constant over time—i.e., intermediaries are long cash bonds and short CDS—but that intermediaries' constraints may sometimes be slack. However, for the CIP deviations, the findings in Du et al. (2018)—i.e., the fact that CIP deviations appear to be correlated with short rate differentials across currencies—suggest that it may be useful to relax A1, thus allowing intermediary positions to switch sign over time. Ultimately, however, it is an empirical question whether A1 and A2 hold in a given setting.

[16]The floating-rate payments on a *zero-coupon swap* are made periodically over time, but there is just a single pre-determined fixed-rate payment at the maturity of the swap. A *zero-coupon swap spread* is the difference between this zero-coupon swap yield and the zero-coupon Treasury yield with the same maturity.

intermediaries must accommodate non-zero end-user demand for the perpetual swap. To close the model in a simple way, we take the limit as net end-user demand for each of these zero-coupon swaps goes to zero from above—i.e., we consider the case where end-users have a positive, but infinitesimal demand to receive fixed on all zero-coupon swaps. As a result, this extension does not change intermediaries' portfolios, so our prior results for the perpetual swap spread (s_t) are unchanged by construction. For simplicity, we further assume that swap spread trades of all maturities consume the same amount of intermediary capital per unit of notional exposure.

Using this approach, we can derive the full term structure of swap spreads $\{s_t^{(n)}\}_{n=1}^N$ from intermediaries' first-order conditions. Specifically, we generalize (15) to obtain the following expression for n-period swap spreads:

$$s_t^{(n)} = \overbrace{n^{-1}\sum_{k=0}^{n-1} \mathrm{E}_t[m_{t+k}]}^{\text{Expected short-rate differentials}} + \overbrace{n^{-1}\sum_{k=0}^{n-1} \mathrm{E}_t[(-\kappa_x)\,\psi\left(\left|d_{t+k}\right|, w_{t+k}\right)]}^{\text{Expected compensation for using scarce capital}} \\ + \overbrace{n^{-1}\sum_{k=0}^{n-1}(-\alpha)\,\mathrm{E}_t[\mathcal{C}_{t+k}^{(n-k)} d_{t+k}]}^{\text{Expected compensation for risk}}, \tag{25}$$

where $r_{t+1}^{s(n)} \equiv n s_t^{(n)} - (n-1)\, s_{t+1}^{(n-1)} - m_t$ is the excess return on an n-period swap spread trade and $\mathcal{C}_t^{(n)} \equiv \mathrm{Cov}_t[r_{t+1}^{s(n)}, r_{t+1}^{s}] = (n-1)\,\frac{\delta}{1-\delta}\mathrm{Cov}_t[s_{t+1}^{(n-1)}, s_{t+1}]$ is the covariance between the returns on the n-period spread trade and the perpetual spread trade. The n-period swap spread equals the expected short-rate differential, the expected compensation for consuming scarce intermediary capital, and the expected compensation for bearing swap spread risk, each averaged over the next n periods. Relative to (15), the only substantive change is that the compensation for risk now depends on how the returns on this zero-coupon swap will *covary* with those on perpetual swaps over time.

Spreads on 1-period swaps are an interesting and important special case. In particular, since 1-period swaps are *riskless* contracts, we have $\mathcal{C}_t^{(1)} = 0$ and the 1-period swap spread depends only the current short rate differential and the current shadow value of intermediary capital:

$$s_t^{(1)} = m_t - \kappa_x \psi\left(\left|d_t\right|, w_t\right). \tag{26}$$

Combining (25) and (26), we can write the n-period spread as the expected 1-period spread averaged over time plus a term reflecting compensation for risk:

$$s_t^{(n)} = \overbrace{n^{-1}\sum_{k=0}^{n-1} \mathrm{E}_t[s_{t+k}^{(1)}]}^{\text{Future expected short-term spreads}} + \overbrace{n^{-1}\sum_{k=0}^{n-1}(-\alpha)\,\mathrm{E}_t[\mathcal{C}_t^{(n-k)} d_{t+k}]}^{\text{Expected compensation for risk}}. \tag{27}$$

Furthermore, under assumptions A1, A2, and A3 (if necessary), n-period spreads take the affine form

$$s_t^{(n)} = A_0^{(n)} + A_m^{(n)} z_t^m + A_d^{(n)} z_t^d + A_w^{(n)} z_t^w, \tag{28}$$

that we characterize in Internet Appendix B.4. Using the resulting affine model, we derive several properties of the term structure of swap spreads in Internet Appendix B.4. We provide the following characterization of the "global" slope of the swap spread curve, defined as the difference between the spread on the perpetual swap (s_t) and the spread on 1-period swaps ($s_t^{(1)}$).

Proposition 5 *The* global *slope of the swap spread curve is given by*

$$s_t - s_t^{(1)} = (A_0 - A_0^{(1)}) + (A_m - A_m^{(1)})z_t^m + (A_d - A_d^{(1)})z_t^d + (A_w - A_w^{(1)})z_t^w. \tag{29}$$

We have $A_0 - A_0^{(1)} = E[s_t - s_t^{(1)}] = -\alpha V \, E[d_t] < 0$; i.e., the term structure of swap spreads slopes downward on average. Furthermore, $A_m - A_m^{(1)} < 0$ and $A_w - A_w^{(1)} < 0$, so increases in short-rate differentials and intermediary wealth each make the spread curve more downward-sloping. Finally, $A_d - A_d^{(1)} < 0$ as long as swap spread risk V is sufficiently large:

$$\frac{\delta}{1-\delta}(1-\rho_d)\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 < V. \tag{30}$$

Proposition 5 states that, when end-user demand to receive the fixed swap rate is positive ($\mathrm{E}[d_t] > 0$), long-maturity swap spreads are lower than short-maturity spreads on average. This downward-sloping spread curve reflects the average compensation for risk that swap intermediaries earn for supplying receive-fixed swaps ($\mathrm{E}[s_t - s_t^{(1)}] = -\alpha V \, \mathrm{E}[d_t] < 0$). Thus, when $\overline{m} > 0$ as in the case of LIBOR-based swaps, the model allows for positive average spreads on short-maturity swaps and negative average spreads on long-maturity swaps.

Proposition 5 says that shocks to both short-rate differentials (m_t) and intermediary wealth (w_t) push up short-dated spreads more than long-dated spreads. The intuition follows from the expectations-hypothesis style logic of (27). Specifically, shocks to m_t and w_t largely impact longer-dated swap spreads by shifting the expected path of 1-period swap spreads. (Indeed, when demand is inelastic ($\gamma = 0$), shocks to m_t and w_t *only* shift the first term in (27).) Since both shocks are mean-reverting, they are expected to have a temporary impact on 1-period spreads. As a result, these shocks have a larger impact on short-dated spreads than on longer-dated spreads. By contrast, demand shocks necessarily impact both terms in (27). However, so long as swap spread risk is sufficiently high, demand shocks primarily impact long-dated swap spreads by shifting the expected compensation for bearing swap spread risk, implying that demand shocks have a greater impact on long-dated spreads.

Because the slope of the swap spread curve depends on the compensation for bearing swap spread risk, it has predictive power for the term structure of returns on swap spread positions. Specifically, we obtain the following result, which is the swap spread analog of the Fama and Bliss (1987) and Campbell and Shiller (1991) result that long-term bonds are expected to outperform short-term bonds when the yield curve is steep:

Proposition 6 *If swap spread risk V is sufficiently large, a larger slope of the swap spread curve predicts that longer-maturity swap spread trades will outperform short-maturity trades. Specifically, the slope coefficient from the forecasting regression*

$$r_{t+1}^s - r_{t+1}^{s(1)} = a + b \cdot (s_t - s_t^{(1)}) + e_{t+1}, \tag{31}$$

satisfies $b > 0$ *so long as (30) holds.*

Intuitively, when V is large, both $\mathrm{E}_t[r^s_{t+1} - r^{s(1)}_{t+1}]$ and $s_t - s^{(1)}_t$ are decreasing in end-user demand d_t. Indeed, $\mathrm{E}_t[r^s_{t+1} - r^{s(1)}_{t+1}]$ is always decreasing in d_t since $\mathrm{E}_t[r^s_{t+1} - r^{s(1)}_{t+1}] = -\alpha V d_t$. As a result, there is a positive time-series association between $s_t - s^{(1)}_t$ and $\mathrm{E}_t[r^s_{t+1} - r^{s(1)}_{t+1}]$. Furthermore, this forecasting relationship disappears if d_t and, hence, the compensation for risk are constant over time—i.e., if $\gamma = \sigma^2_d = 0$.

Internet Appendix B.4 provides a range of additional results on the term structure of swap spreads. In particular, we can also characterize the local behavior of the swap spread curve—i.e., how $s^{(n)}_t - s^{(n-1)}_t$ behaves as a function of n—in the limit where end-user demand is completely inelastic ($\gamma = 0$). We find that $A^{(n)}_0$ is decreasing in maturity n—i.e., on average, the swap spread curve is locally downward-sloping at all maturities. When $\overline{m} = E[m_t] \geq 0$ is sufficiently small, $A^{(n)}_0 < 0$ for all n. When $\overline{m}$ is sufficiently large, $A^{(n)}_0 > 0$ for all n. In the intermediate case where $\overline{m}$ is moderately positive, $A^{(n)}_0 > 0$ for small n and $A^{(n)}_0 < 0$ for larger n. $A^{(n)}_m$ reflects variation in the expected future short rate differentials (m_t) over the life of the swap and, thus, is positive and locally downward-sloping across maturities n. $A^{(n)}_w$ reflects variation in the impact of intermediary wealth on expected future balance sheet costs ($-\kappa_x \psi_t$) over the life of the swap and, thus, is positive and locally downward-sloping across maturities n. $A^{(n)}_d$ is negative for all maturities n and reflects both the (i) expected balance sheet costs and (ii) compensation for risk over the life of the swap. When the volatility of swap spreads is sufficiently low, (i) dominates and $A^{(n)}_d$ is an increasing function of maturity n. When the volatility of swap spreads is higher, (ii) dominates. In this case, $A^{(n)}_d$ is downward-sloping across maturities n when ρ_d is sufficiently high and is a U-shaped function of maturity n when ρ_d is lower. By continuity, this characterization remains valid so long as the demand elasticity γ is not too large.

4 Evidence

4.1 Data and measurement

We use two main types of time-series data in our empirical analysis: data on swap spreads and data that we use to proxy for intermediaries' net positions in the swap spread trade. Our main dataset is weekly and runs from July 2001 to June 2020.

Swap spreads. We obtain fixed rates for plain-vanilla LIBOR swaps, OIS swaps, and SOFR swaps from Bloomberg and constant maturity Treasury yields from the Federal Reserve's H.15 Statistical Release. Data availability varies by swap product and maturity. For instance, we have data on 30-year LIBOR swap rates over our full sample, but data on 30-year OIS and SOFR swap rates only become available in September 2011 and December 2018, respectively. We compute swap spreads as the difference between the fixed swap rate and the constant maturity Treasury yield with the same maturity. We obtain 3-month LIBOR, the 3-month OIS rate, and the rate on 3-month Treasury general collateral repurchase agreements (GC repo) from Bloomberg.

Following Boyarchenko et al. (2020), we compute the h-week cumulative returns on the spread trade involving n-year LIBOR swaps as:

$$r^{s(n)}_{t\to t+h} = \sum_{k=0}^{h-1} [\overbrace{\frac{1}{52}(s^{(n)}_{t+k} - (i^{LIBOR}_{t+k} - i^{REPO}_{t+k}))}^{\text{Carry}} - \overbrace{DV01^{(n)}_{t+k} \times (s^{(n)}_{t+k+1} - s^{(n)}_{t+k})}^{\text{Mark-to-market loss}}], \tag{32}$$

where $s^{(n)}_t$ in the n-year LIBOR swap spread, $i^{LIBOR}_t - i^{REPO}_t$ is the difference between 3-month LIBOR and GC repo, and $DV01^{(n)}_t$ is the dollar value of a basis point for this trade—i.e., the sensitivity of the position's mark-to-market value to changes in the swap spread.[17] The term in square brackets in (32) is the 1-week spread trade return from week $t+k$ to $t+k+1$. As in (3), this weekly return is the sum of a carry component known at $t+k$ and a mark-to-market component that depends on the change in swap spreads from $t+k$ to $t+k+1$. The h-week cumulative return from t to $t+h$, $r^{s(n)}_{t\to t+h}$, is the sum of the next h weekly returns after time t. We focus on $h = 13$-week and $h = 52$-week returns in our empirical tests—i.e., 3- and 12-month returns.

We primarily focus on the 30-year LIBOR swap spread. The 30-year swap spread corresponds closely to the long-term arbitrage spread in our model. Intermediaries who take spread positions using 30-year swaps both consume scarce capital and face the risk of suffering significant short-term mark-to-market losses if spreads move against them. Moreover, 30-year LIBOR swaps are a major source of duration for end users (Klingler and Sundaresan, 2019). However, we also examine shorter-dated swap spreads and the associated spread trade returns.

Intermediaries' swap spread positions. Data on intermediaries' positions in the swap spread trade are not directly available. However, we can use primary dealers' net position in Treasury securities to proxy for the scale of their net swap spread trade position. Broker-dealers are critical intermediaries in the swap market (TBAC, 2021). However, dealers have minimal net exposure to changes in the overall level of long-term interest rates.[18] In particular, while dealers accommodate net end-user demand to either receive or pay the fixed swap rate, they typically take offsetting positions in Treasuries to hedge the resulting interest-rate risk. Thus, primary dealers' net position in Treasuries—especially those maturing in more than one year—is a hedge that partially mirrors their net position in receive-fixed interest rate swaps.

To construct our measure of primary dealers' net position, we download weekly data on primary dealers' aggregate positions in Treasury securities from the Federal Reserve Bank of New York's website. These weekly aggregates are derived from Form A of the FR2004 reports (the Primary Government Securities Dealers Reports) that primary dealers are required to file with the Federal Reserve and which contain detailed dealers' positions as of market-close each Wednesday. Form FR2004 underwent a significant revision in July 2001, which is why our main sample begins at that time. Dealers' net Treasury position, denoted *PD-UST-Net*$_t$, is calculated as the difference

[17] The DV01 for a n-year swap spread trade is the n-year annuity factor (divided by $10,000$). Assuming the fixed swap leg is paid semi-annually, $DV01^{(n)}_t \equiv (1/2)\sum_{i=1}^{2n}(1+Y^{(n)}_t/2)^{-i}$ where $Y^{(n)}_t$ is the n-year swap yield. Technically, the $DV01$s for the Treasury and swap legs of the spread trade will differ slightly because the Treasury and swap yields differ. Accounting for this small discrepancy has almost no impact on our results, so we have chosen to use the simpler definition of returns in (32).

[18] For instance, O'Brien and Berkowitz (2007) show that the variation in dealers' trading revenues from interest-rate risk exposure is small relative to the variation in total trading revenues, which also includes fees and spreads.

between the market values of dealers' long and short Treasury positions. We focus on nominal coupon-bearing Treasury securities—which are most likely to be tied to swap spread positions—and exclude Treasury bills, TIPS, and FRNs. We will sometimes write $\hat{x}_t = -1 \times \textit{PD-UST-Net}_t$ to emphasize that $-1 \times \textit{PD-UST-Net}_t$ is our proxy for the scale of dealers' net position in the receive-fixed swap spread trade. By market clearing, this means that $\hat{d}_t = \textit{PD-UST-Net}_t$ is our proxy for end-user's net demand to receive the fixed swap rate.

Consistent with the logic of our measurement approach, primary dealers' net Treasury positions do not appear to be tightly linked to their net exposure to interest-rate risk. To demonstrate this, we obtain the value-weighted (by book equity) average of primary dealers' interest-rate Value-at-Risk from Anderson and Liu (2021). The correlation between weekly changes in primary dealers' net Treasury position and changes in their interest-rate Value-at-Risk is just 0.05 between 2001 and 2018 and 0.08 between 2009 and 2018. The fact that there is little relationship between dealers' net Treasury position and their interest-rate Value-at-Risk suggests that these Treasury positions largely hedge interest-rate exposures stemming from their positions in interest rate swaps and other fixed-income instruments.

That said, we recognize that $\textit{PD-UST-Net}_t$ is a noisy proxy for the scale of dealers' swap spread positions; a variety of other factors also affect dealers' net Treasury position. For instance, primary dealers play a role in intermediating Treasury auctions, absorbing some portion of new Treasury issuance into short-term inventory and distributing these securities to end investors over time (Fleming and Rosenberg, 2008). In recent years, the effective quantity of Treasury issuance that dealers must intermediate has also been impacted by the Federal Reserve's Large-Scale Asset Purchase policies (i.e., Quantitative Easing), which are regularly implemented by adjusting the amount that the Federal Reserve purchases at auctions. Furthermore, dealers use Treasuries to hedge their inventories of a variety of other fixed-income instruments beyond swaps, especially their corporate bonds inventories. However, since we typically use $\textit{PD-UST-Net}_t$ as an independent variable in our regression specifications, measurement error will generally attenuate our findings, biasing us against finding a significant relationship between $\textit{PD-UST-Net}_t$ and swap spreads or swap spread trade returns. Indeed, our results are strongest when we control for other factors—e.g., the gross quantity of Treasury issuance—that are unrelated to the scale of dealers' swap spread positions but that independently affect $\textit{PD-UST-Net}_t$.

Additional data. In addition to data on their Treasury positions, we collect data on primary dealers' net positions in corporate debt securities from the FR2004 reports. From the FR2004, we also collect the amount of Treasuries collateralizing financing agreements in which a primary dealer is the cash lender—i.e., dealers' Treasury "securities in" —or the cash borrower—i.e., dealers' Treasury "securities out." These financing arrangements include reverse repos, securities borrowing, and securities received as margin collateral. Through these arrangements primary dealers play a key role in allowing other investors—notably hedge funds—to finance their Treasury positions (He et al., 2022). We use the amount of cash primary dealers lend against Treasuries as a noisy proxy for the scale of hedge funds' swap spread positions. And, we use the difference between primary dealers' "securities out" and their "securities in" as an alternative proxy for their levered exposure to Treasuries. We also obtain data on monthly Treasury issuance from SIFMA and weekly data on the Federal Reserve's Treasury holdings from the H.4.1 release.

We gather data on a variety of additional market prices. From Bloomberg, we obtain the spread on 10-year U.S. Treasury credit default swaps (CDS), the VIX SP500 option-implied equity volatility index, and the VXTY option-implied 10-year Treasury futures volatility index. We download the Adrian et al. (2013) estimate of the term premium component of 10-year Treasury yields from the Federal Reserve Bank of New York's website. We obtain Hu et al. (2013)'s measure of yield curve noise—the root mean-squared yield error from fitting a Svensson (1994) model to the cross-section of Treasuries on each date—from Jun Pan's website.

Finally, we collect proxies for several factors that are thought to shift end users' demand for or intermediaries' supply of swaps. First, we obtain the modified duration of the Barclays U.S. MBS index from Datastream. The duration of outstanding MBS is a proxy for relative-value mortgage investors' desire to pay the fixed swap rate. However, as discussed in Section 4.5, MBS duration may also proxy for a variety of related "convexity-driven" hedging motives. Second, we compute Klingler and Sundaresan (2019)'s measure of aggregate defined-benefit pension underfunding, which affects pensions' demand to receive the fixed rate on long-dated swaps. Specifically, using quarterly data from the Financial Accounts of the United States, we compute $\textit{Pension-UFR}_t = (\textit{Pension-Sponsor-Claims}_t / \textit{Pension-Asset}_t)$ where $\textit{Pension-Sponsor-Claims}_t$ is the total claims of private and public defined-benefit pensions on their sponsors and $\textit{Pension-Asset}_t$ is total defined-benefit pension assets. Third, we obtain the monthly gross issuance of domestic public bonds by U.S. non-financial corporations from the Federal Reserve Board, which is a proxy for non-financial corporations' demand to receive the fixed rate. Finally, as a proxy for intermediaries' wealth, we obtain primary dealers' equity capital ratio—the ratio of market equity to the sum of market equity and book debt—from He et al. (2017).

4.2 The emergence of negative swap spreads

Consistent with the logic of the general model in Section 3.2.1, Panel A of Figure 1 shows that 30-year LIBOR swap spreads and primary dealers' net Treasury positions simultaneously switched signs in early 2009. Specifically, (15) says that, absent a large short-rate differential (m_t), swap spreads should have the opposite sign of $\hat{d}_t = -\hat{x}_t = \textit{PD-UST-Net}_t$. Indeed, 30-year LIBOR swap spreads averaged 45 basis points from 2001 to 2008 and primary dealers' net short position in Treasuries averaged \$96 billion. By contrast, 30-year LIBOR swap spreads have been consistently negative since 2009, averaging -25 basis points from 2009 to 2020. And, during this 2009–2020 period primary dealers' net long position in Treasuries averaged \$59 billion. In contemporaneous work, Du et al. (2022) also document that primary dealers' net Treasury position and swap spreads move inversely over time and they emphasize this regime change in early 2009.

Panel B of Figure 1 shows the evolution of the 30-year OIS and SOFR swap spreads. These two spreads are highly correlated with the 30-year LIBOR swap spread, but are even more negative on average because the underlying short-rate differentials are negligible. Almost by definition, a non-zero SOFR swap spread represents a violation of LoOP.

Panel C of Figure 1 shows that the term structure of swap spreads changed considerably in

2009. While the term structure of spreads was fairly flat before the GFC, it has become steeply downward-sloping since 2009. Specifically, the average difference between 30-year and 10-year spreads was negligible prior to 2009 but has averaged -29 basis points between 2009 and 2020, with 30-year swap spreads being consistently more negative than 10-year spreads.

This post-2008 combination of negative long-term swap spreads, a downward-sloping swap spread curve, and dealers' net long position in Treasuries is consistent with our models' predictions in a regime where leverage-constrained and risk-averse intermediaries must enter into a pay-fixed swap spread trade to accommodate net end-user demand to receive fixed. Thus, in what follows, we focus on the post-2008 period, examining the specific predictions that pertain to this regime.

4.3 Swap spread risk

We begin by testing our model's prediction that intermediaries' exposure to swap-spread risk shapes the equilibrium expected returns on the swap spread trade. Specifically, Proposition 2 says that $\mathrm{E}_t[r^s_{t+1}] = -\kappa_x \cdot \psi_t - \alpha V \cdot d_t$ —i.e., $\mathrm{E}_t[r^s_{t+1}]$ consists of two terms. First, $\mathrm{E}_t[r^s_{t+1}]$ is shaped by the cost of consuming scarce intermediary capital ($-\kappa_x \cdot \psi_t$), which obtains even if the trade is riskless. Second, $\mathrm{E}_t[r^s_{t+1}]$ depends on the required compensation for bearing mark-to-market risk ($-\alpha V \cdot d_t$), which is proportional to the scale of intermediaries' net position in the swap spread trade (d_t). Thus, our theory predicts that the scale of intermediaries' swap spread positions should predict returns even when controlling for the shadow value of intermediary capital (ψ_t).

We test this implication by running predictive regressions that relate the future h-week returns on the receive-fixed 30-year swap spread trade $r^{s(30)}_{t\to t+h}$ on primary dealers' current net Treasury position and the 3-month OIS swap spread (*ST-OIS-Spread*$_t$), which we use as a proxy for the current shadow value of intermediary capital:[19]

$$r^{s(30)}_{t\to t+h} = \alpha + \beta_1 \cdot \textit{PD-UST-Net}_t + \beta_2 \cdot \textit{PD-UST-Credit}_t + \gamma_1 \cdot \textit{ST-OIS-Spread}_t + \boldsymbol{\gamma}_2' \mathbf{x}_t + \epsilon_{t\to t+h}. \quad (33)$$

We estimate these specifications for $h = 13$- and 52-week returns (i.e., 3- and 12-month returns).[20] These regressions also include controls for factors that arguably have an independent effect on primary dealers' inventory of Treasuries, namely the 12-month moving average of monthly gross Treasury issuance, the Federal Reserve's Treasury holdings, and primary dealers' net position in corporate debt securities; see Fleming and Rosenberg (2008). Changes in these factors may alter *PD-UST-Net*$_t$ even holding fixed the scale of dealers swap spread positions, making it a noisy proxy for dealers' swap spread positions. Thus, the purpose of these controls is to isolate the residual component of *PD-UST-Net*$_t$ that is most informative about dealers' swap spread positions.

As can be seen by comparing Panels A and B of Table 1, without controls, dealers' net position strongly predicts swap spread returns at both 3-month and 12-month horizons in the sample from

[19]This choice is motivated by the model. From (26), the shadow value is proportional to the negative of the short-dated SOFR swap spread—i.e., the swap spread with no interest-rate differential. Because data on SOFR swaps is only available for a small part of our sample, we use the closely related short-dated OIS spread.

[20]Since these regressions use overlapping h-week returns, the residuals will be serially correlated. To draw proper inferences, we compute Newey and West (1987) standard errors allowing for serial correlation at up to $\lceil 1.5 \times h \rceil$ weeks. We assess significance using the fixed-b asymptotic theory of Kiefer and Vogelsang (2005) which yields more conservative p-values and has better finite-sample properties than traditional Gaussian asymptotics.

2009m1 to 2018m6, but not when the sample is extended to 2020m6. From mid-2018 until the end of 2019 Treasury issuance increased due to the Tax Cuts and Jobs Act of 2017 and the Federal Reserve started to reduce the Treasury holdings it had accumulated through the Quantitative Easing policies it had pursued since 2009. And 2020 then witnessed a surge in Treasury issuance due to the vast scale of the fiscal policy response to the COVID-19 pandemic and a resumption of large-scale Treasury purchases by the Fed. Adding controls that aim to capture these factors restores the significance of primary dealers' net position in the sample from 2009m1 to 2020m6, and also increases its significance in the sample from 2009m1 to 2018m6.

Recalling that *PD-UST-Net*$_t = -\hat{x}_t = \hat{d}_t > 0$ in the post-GFC regime, the negative coefficient on *PD-UST-Net*$_t$ means that dealers earn larger expected returns on their pay-fixed spread trades when they have greater exposure to this trade. Thus, the results in Table 1 align with the predictions from Proposition 2. The R^2 of the predictive regression for 3-month returns without additional controls reaches 13.2% in the sample from 2009m1 to 2018m6.

Institutions other than primary dealers—particularly, fixed-income hedge funds—also function as intermediaries in the swap market and engage in swap spread trades. This fact motivates the addition of a second proxy for intermediaries' swap positions to our forecasting regression: the quantity of short-term credit collateralized by Treasuries that primary dealers extend to other institutions referred to as "Treasury securities in" and denoted by *PD-UST-Credit*$_t$. The idea is that these financing arrangements, in part, reflect the swap spread trades of hedge funds. This second proxy also predicts swap spread returns, albeit less strongly than primary dealers' own position.

Focusing on the sample from 2009m1 to 2018m6, Panel A of Table 2 shows that primary dealers' net position in long-dated Treasuries—those maturing in 11 or more years—predicts the returns to the 30-year swap spread trade ($PD\text{-}UST^{11+}\text{-}Net_t$). We also obtain similar results using a measure of dealers' levered Treasury exposure, $PD\text{-}UST\text{-}Lev_t$, computed as the difference between dealers' cash borrowing through repurchase agreements ("Treasury securities out") and their cash lending through repo ("Treasury securities in").[21] Panel B of Table 2 shows that primary dealers' net Treasury position also predicts the returns to the 10-year swap spread trade. However, when we forecast the returns on the 10-year trade, our forecasting variables attract coefficients that are smaller in absolute magnitude and the statistical significance is not quite as strong as when forecasting returns on the 30-year trade. This diminished predictive power is consistent with the premise that the 30-year swap spread trade is subject to greater swap spread risk.

4.4 Term structure

Our analysis thus far has primarily focused on the long end of the swap market. We have focused on long-dated swaps for several reasons. First, long-dated swap spread trades are subject to greater risk and, as a result, long-maturity swap spreads are more informative about the swap spread risk channel that we emphasize. Second, short-dated and long-dated swap spreads may be driven by

[21] A large fraction of dealers' repo cash borrowing corresponds to matched-book repo transactions where the dealer is intermediating between a cash borrower like a hedge fund and a cash lender like a money market fund. Thus, we use the difference between cash borrowing and cash lending as a proxy for dealers' levered Treasury exposure. The correlation between the resulting measure, $PD\text{-}UST\text{-}Lev_t$, and our baseline $PD\text{-}UST\text{-}Net_t$ variable is 0.82.

different factors in a way that is not captured by our simple three-factor term structure model. For instance, on the demand side, end-user demand for short-dated swaps may be driven by a different set of factors from those that drive demand for long-dated swaps. (We pursue an extension along these lines below.) Relatedly, on the supply side, intermediaries face a number of different balance sheet constraints, each with its own potentially time-varying shadow value (Siriwardane et al., 2021). And, while some balance sheet constraints like the Supplemental Leverage Ratio apply equally to all swap spread positions irrespective of their maturity and risk, other balance sheet constraints—e.g., margin requirements—are typically less stringent for shorter-dated maturities (Boyarchenko et al., 2020). In other words, in reality, there may be multiple demand factors and multiple supply constraints, each playing a more or less important role at different points along the curve. That said, it is instructive to ask whether the term structure of spreads is consistent with the predictions of our simple term structure model.

The average term structure of swap spreads. Figure 2 shows the average term structures of LIBOR, OIS, and SOFR spreads, starting from 2009 or the date for which data became available. All three term structures are monotonically downward-sloping on average. Further, for all maturities, average OIS and SOFR swap spreads are lower than like-maturity LIBOR swap spreads. In particular, short-dated LIBOR swap spreads are positive on average, whereas short-dated OIS and SOFR swap spreads are negative on average.

These patterns fit well with Propositions 1 and 5. Intermediaries require compensation for committing their scarce capital to positions in the pay-fixed swap spread trade and for bearing the risk associated with these positions. Short-date differentials for OIS and SOFR swaps are negligible, so our model predicts negative average OIS and SOFR swap spreads for all maturities. And, since longer maturity swaps are associated with greater convergence risk and a greater expected return compensation, the term structure of OIS and SOFR swap spread should slope downward on average. Relative to OIS and SOFR swaps, the difference between LIBOR and repo rates pushes up the average level of LIBOR swap spreads. Thus, we would expect to see positive spreads on short-dated LIBOR swaps and negative spreads on long-dated LIBOR swaps.

Fama-Bliss style forecasting regressions. Assuming that intermediaries face significant swap spread risk, Proposition 6 says that long-dated swap spread trades are expected to outperform short-dated trades when the swap spread curve is steep. Intuitively, greater end-user demand to receive fixed pushes down the expected returns on long-dated spread trades relative to those on short-dated spread trades—i.e., $\mathrm{E}_t[r^s_{t+1} - r^{s(1)}_{t+1}] = -\alpha V d_t$. And, assuming that spread risk is sufficiently large, the slope of the spread curve $(s_t - s^{(1)}_t)$ is also decreasing in end-user demand, giving rise to a positive relationship between $\mathrm{E}_t[r^s_{t+1} - r^{s(1)}_{t+1}]$ and $(s_t - s^{(1)}_t)$. In this way, Proposition 6 allows us to further highlight the importance of swap spread risk.

We test Proposition 6 using an analogue of the regression specification from Fama and Bliss (1987) applied to swap spread trades. Specifically, for two maturities $n > n'$, we estimate

$$r^{s(n)}_{t\to t+h} - r^{s(n')}_{t\to t+h} = \alpha + \beta \cdot (s^{(n)}_t - s^{(n')}_t) + \epsilon_{t\to t+h}, \qquad (34)$$

and our model predicts that $\beta > 0$. As reported in Table 3, the slope of the swap spread curve is a significant predictor of excess returns on longer-dated swap spread trades over shorter-dated trades in both the 2009m1 to 2020m6 sample (Panel A) and the 2009m1 to 2018m6 sample (Panel B) and for a variety of different maturity pairs $n > n'$. This result supports a risk-based explanation of swap spreads and, in particular, is consistent with our model's prediction that long-term swap spreads should reflect compensation for demand-and-supply imbalance risk.

The impact of supply and demand on the term structure. Our model also suggests that, in absolute terms, supply shocks should have a larger impact on short-dated swaps while demand shocks should have their greatest impact on intermediate- or longer-dated swaps. In Tables D-1 and D-2 of the Internet Appendix, we test this additional prediction by regressing the 3-month change in swap spreads at various maturities n on several proxies for the change in demand and supply. Specifically, we use the same set of proxy variables introduced in Section 4.5 below. We find that changes in demand generally have their largest impact on longer-dated spreads and, thus, alter differences between longer- and shorter-dated swap spreads in the way predicted by our theory. However, the results for our supply proxies are less clear cut.

4.5 Demand and supply decomposition

Having validated *PD-UST-Net*$_t$ as a useful proxy for intermediaries' net position in the receive-fixed swap spread trade, we now use a structural vector auto-regression (SVAR) to disentangle the effects of end-user demand and intermediary supply on the level of 30-year swap spreads. Leveraging Proposition 4, we set-identify our structural VAR using the sign restrictions approach of Uhlig (2005). We then estimate the latent underlying demand and supply factors and decompose swap spreads into contributions from these two factors. We shed light on the primitive factors that drive swap spreads by examining the time-series correlates of our estimated demand and supply factors. Finally, we look at the respective roles of end-user demand and intermediary supply in shaping the expected returns to swap arbitrage. To avoid the need to introduce exogenous controls in our SVAR, for simplicity, we focus on the 2009m1 to 2018m6 sample in this exercise.

Estimation approach. We begin with the reduced-form representation of the VAR implied by our structural model. Specifically, we consider a bivariate VAR involving 30-year swap spreads ($s_t^{(30)}$) and end-users' net position in receive-fixed swaps proxied using primary dealers' net position in Treasuries ($\hat{d}_t = -\hat{x}_t =$ *PD-UST-Net*$_t$):

$$\overbrace{\begin{bmatrix} s_t^{(30)} \\ \hat{d}_t \end{bmatrix}}^{\mathbf{y}_t} = \mathbf{c} + \sum_{l=1}^{L} \mathbf{C}_l \overbrace{\begin{bmatrix} s_{t-l}^{(30)} \\ \hat{d}_{t-l} \end{bmatrix}}^{\mathbf{y}_{t-l}} + \overbrace{\begin{bmatrix} \xi_t^{s^{(30)}} \\ \xi_t^{\hat{d}} \end{bmatrix}}^{\boldsymbol{\xi}_t}. \tag{35}$$

Equation (35) corresponds to our structural model assuming (i) the short-rate differential is always zero (i.e., $m_t = 0$) and (ii) allowing for a more general lag structure in the autoregression—i.e., our model assumes $L = 1$ for simplicity. Naturally, the shocks $\boldsymbol{\xi}_t$ in this reduced-form VAR each reflect a combination of structural shocks to both end-user demand and intermediary supply and, thus, do not have clear-cut economic interpretations.

For parsimony, we omit the short-rate differential m_t from (35). As discussed below, adding m_t to the VAR has minimal impact on our results. We select a lag length of $L = 2$ weeks in (35) by minimizing the Akaike Information Criterion (AIC). Our results are robust to alternate choices of L as well as to the inclusion of a deterministic time trend.

Following Proposition 4, we posit that the reduced-form VAR in (35) results from (i) a mapping from a set of unobserved structural demand and supply factors to these observable equilibrium outcomes and (ii) a dynamic law of motion for the structural factors. Specifically, the assumed mapping from latent structural factors to observable outcomes is

$$\overbrace{\begin{bmatrix} s_t^{(30)} \\ \hat{d}_t \end{bmatrix}}^{\mathbf{y}_t} = \mathbf{a} + \mathbf{A} \overbrace{\begin{bmatrix} z_t^d \\ z_t^w \end{bmatrix}}^{\mathbf{z}_t}. \tag{36}$$

And, the assumed law of motion for the structural factors is

$$\overbrace{\begin{bmatrix} z_t^d \\ z_t^w \end{bmatrix}}^{\mathbf{z}_t} = \sum_{l=1}^{L} \mathbf{D}_l \overbrace{\begin{bmatrix} z_{t-l}^d \\ z_{t-l}^w \end{bmatrix}}^{\mathbf{z}_{t-l}} + \overbrace{\begin{bmatrix} \varepsilon_t^d \\ \varepsilon_t^w \end{bmatrix}}^{\boldsymbol{\varepsilon}_t}, \tag{37}$$

where $\boldsymbol{\varepsilon}_t$ is a vector of orthogonal structural shocks that each have unit variance—i.e., $\text{Var}_{t-1}[\boldsymbol{\varepsilon}_t] = \text{Var}[\boldsymbol{\varepsilon}_t] = \mathbf{I}$.[22] Combining (36) and (37), we obtain a bivariate structural VAR that we can take to the data:

$$\mathbf{y}_t = \left(\mathbf{I} - \sum_{l=1}^{L} \mathbf{A}\mathbf{D}_l\mathbf{A}^{-1}\right)\mathbf{a} + \sum_{l=1}^{L} (\mathbf{A}\mathbf{D}_l\mathbf{A}^{-1})\mathbf{y}_{t-l} + \mathbf{A}\boldsymbol{\varepsilon}_t, \tag{38}$$

which corresponds to (22) in the model.

We set identify this structural VAR by imposing the sign restrictions from Proposition 4 on the $\mathbf{A}$ matrix that maps structural shocks $\boldsymbol{\varepsilon}_t$ into reduced-form shocks $\boldsymbol{\xi}_t$. Specifically, we assume the four elements of $\mathbf{A}$ satisfy the following sign restrictions:

$$\overbrace{\begin{bmatrix} \xi_t^{s(30)} \\ \xi_t^{\hat{d}} \end{bmatrix}}^{\boldsymbol{\xi}_t} = \overbrace{\begin{bmatrix} - & + \\ + & + \end{bmatrix}}^{\mathbf{A}} \overbrace{\begin{bmatrix} \varepsilon_t^d \\ \varepsilon_t^w \end{bmatrix}}^{\boldsymbol{\varepsilon}_t}. \tag{39}$$

Intuitively, if there are positive (negative) reduced-form innovations to both $|s_t^{(30)}|$ and $\hat{d}_t$, there must have been a positive (negative) shock to end-user demand (ε_t^d). Conversely, if there is negative (positive) innovation to $|s_t^{(30)}|$ and a positive (negative) innovation to $\hat{d}_t$, there must have been a positive (negative) shock to intermediary supply (ε_t^w).[23]

We set identify $\mathbf{A}$ using the pure sign restrictions approach of Uhlig (2005). Since $\text{Var}[\boldsymbol{\varepsilon}_t] = \mathbf{I}$

[22]The assumption that the demand and supply shocks are orthogonal is necessary for our identification approach and is useful for these shocks to have a straightforward interpretation. While imperfect, the assumption that demand and supply shocks are orthogonal is reasonable for the swap market because the institutions that typically receive the fixed rate—end users such as pension funds—and those that pay the fixed rate—intermediaries such as broker-dealers—have very different investment objectives and face different institutional constraints.

[23]These statements follow from inverting (39) and using the sign restrictions on $\mathbf{A}$ to sign the elements of $\mathbf{A}^{-1}$.

—i.e., the structural shocks are assumed to be orthogonal and we are only attempting to identify their impact "up to scale"—the observed variance of reduced-form shocks, Var[$\boldsymbol{\xi}_t$] $= \mathbf{A}\mathbf{A}'$, provides us with three restrictions on the four elements of $\mathbf{A}$. The sign restrictions in (39) do not allow us to point identify $\mathbf{A}$. However, they provide a partial identification, ruling out many candidates that satisfy $\mathbf{A}\mathbf{A}' =$ Var[$\boldsymbol{\xi}_t$] and leaving us with an identified set of $\mathbf{A}$s. Following Fry and Pagan (2005, 2011) and Cieslak and Pang (2021), we select the $\mathbf{A}$ in this set that is closest to the median value in the set. To estimate the latent structural demand and supply factors, we invert (36) using this "closest-to-median" matrix. Letting $\hat{\mathbf{A}}$ denote the closest-to-median value of $\mathbf{A}$, our estimated factors are $\hat{\mathbf{z}}_t = \hat{\mathbf{A}}^{-1}(\mathbf{y}_t - \hat{\mathbf{a}})$, where $\hat{\mathbf{a}} = (\mathbf{I} - \sum_{l=1}^{L} \hat{\mathbf{C}}_l)^{-1}\hat{\mathbf{c}}$.[24]

In Internet Appendix C.2, we also consider a tri-variate VAR which adds the LIBOR-repo spread and corresponds to our model when there is a non-zero short-rate differential ($m_t \neq 0$). The structural shocks in this tri-variate VAR are identified using the sign and zero restrictions implied by Proposition 4. Including the short-rate differential in the VAR has almost no impact on our estimates of the latent demand and supply factors. Furthermore, our estimates suggest that the short-rate differential explains very little of the time-series variation in 30-year swap spread observed since 2009; see Figure D-1 of the Internet Appendix. The finding that the short-rate differential plays only a small role in explaining 30-year swap spreads is consistent with our model that predicts that transient fluctuations in m_t—e.g., due to a temporary rise in concerns about the creditworthiness of large banks—should play only a minor role in explaining movements in long-dated swap spreads.[25] Thus, we use the more parsimonious specification (35) as our baseline.

Estimated structural decomposition. We begin by reporting and interpreting our closest-to-median estimate, $\hat{\mathbf{A}}$. Since we can only identify the structural shocks "up to scale," our estimate $\hat{\mathbf{A}}$ corresponds to $\mathbf{A}\boldsymbol{\Sigma}^{1/2}$ in our model. Our closest-to-median estimate is given by:

$$\hat{\mathbf{A}} = \begin{bmatrix} \sigma_d A_d & \sigma_w A_w \\ \sigma_d (1 + \gamma A_d) & \gamma \sigma_w A_w \end{bmatrix} = \begin{bmatrix} -2.24 & 2.26 \\ 7.07 & 7.04 \end{bmatrix}. \tag{40}$$

Using the estimated coefficients from our structural VAR, we can use our model to back out some interesting economic parameters. First, taking the ratio of the estimated impact of a 1-standard deviation shock to z_t^w on the equilibrium quantity ($\gamma\sigma_w A_w$) to the corresponding impact on prices ($\sigma_w A_w$), we obtain an IV-like estimate of the elasticity of end-user demand (γ). Specifically, our structural VAR estimates suggest that $\gamma = 7.04/2.26 = 3.1$—i.e., each basis point increase in swap spreads raises net end-user demand to receive fixed by $\$3.1$ billion.

Using this estimate of γ and the estimated impact of demand shocks on quantities and spreads, we infer that $\sigma_d = \sigma_d (1 + \gamma A_d) - \gamma \times \sigma_d A_d = 7.07 - 3.1 \times -2.24 = 14$. Thus, at a weekly

[24]This closest-to-median $\mathbf{A}$ corresponds to a particular matrix in the identified set and has a structural interpretation. Working with the closest-to-median $\mathbf{A}$ is preferred to working with the median value of $\mathbf{A}$ in the identified set because the latter lacks a structural interpretation. However, the demand and supply factors implied by the closest-to-median $\mathbf{A}$ are almost identical to the median factors across all $\mathbf{A}$s in the set, thus satisfying the specification check for this partial identification technique suggested by Fry and Pagan (2005, 2011). See Figure 4.

[25]This finding is consistent with the fact that long-dated LIBOR, OIS, and SOFR swap spreads are very highly correlated in both levels and changes. By contrast, spreads on short-dated LIBOR swaps—which are sensitive to movements in the LIBOR-repo spread—are not highly correlated with those on short-dated OIS and SOFR swaps.

horizon, we estimate that a 1-standard deviation shock to end-user demand is $\$14$ billion. This then implies that a $\$1$ billion shock to end-user demand to receive fixed reduces 30-year swap spreads by 0.16 basis points—i.e., $A_d = A_d\sigma_d/\sigma_d = -2.24/14 = -0.16$.

Several caveats are in order. First, dealers are only one set of intermediaries in the swap market, the other important set being hedge funds. Thus, it seems safe to conclude that these dollar estimates are a lower-bound on the relevant market-wide quantities which sum across all relevant swap intermediaries. Second, we have arrived at these interpretations by taking our stylized model and our identification approach rather literally. For instance, our model and identification approach assume that the structural shocks to demand and supply are orthogonal. This assumption is both conventional and clarifying, but there is little reason to think this is a perfect description of reality. Nonetheless, our estimates seem to be quantitatively plausible, lending credence to our approach.

Next, Figure 3 shows our estimated structural decomposition of swap spreads from 2009m1 to 2018m6 as in (36). Specifically, using $\hat{\mathbf{A}}$ and $\hat{\mathbf{z}}_t$, we decompose $s_t^{(30)} - \mathrm{E}[s_t^{(30)}]$ into the sum of contributions from end-user demand ($\hat{A}_d\hat{z}_t^d$) and intermediary supply ($\hat{A}_w\hat{z}_t^w$). Figure 3 shows that end-user demand and intermediary supply both play important roles in explaining the time-series variation in swap spreads. For instance, our decomposition suggests that a large outward shift in intermediary supply reduced the magnitude of (negative) swap spreads starting in late-2012. This outward supply shift is consistent with an easing of financial conditions following the crest of the Eurozone crisis. Then, a large inward shift in supply from late-2014 until late-2015 pushed swap spreads far into negative territory. This estimated inward supply shift coincides with a series of regulatory changes that arguably increased intermediaries' balance-sheet costs, including the finalization of the Supplementary Leverage Ratio in September 2014 and the implementation of the Volker Rule in July 2015 (Boyarchenko et al., 2020). However, we also find that the period of highly negative swap spreads between late 2014 and 2018 cannot be solely attributed to a contraction in intermediary supply. Indeed, from mid-2016, our estimates suggest that rising end-user demand to receive the fixed rate pushed swap spreads even further below zero.

The forecast error variance decomposition from our structural VAR—the estimated contributions of the two structural shocks to swap spread innovations at various horizons—confirms that both demand and supply shocks play an important role in driving spreads. Demand and supply shocks each explain roughly 50% of the innovations to 30-year swap spreads at a weekly horizon. Our finding that both demand and supply shocks play an important role in driving swap spreads contrasts with Goldberg and Nozawa (2021) who find that, in the corporate bond market, intermediary supply shocks play a dominant role in driving variation in market liquidity. Interestingly, our estimates suggest that supply shocks have a more persistent effect on swap spreads compared to demand shocks. And, as a result, the contribution of supply shocks to the innovations to 30-year swap spreads increases to roughly 65% at an annual horizon.

Drivers of demand and supply. An advantage of our identification approach is that we do not need to specify proxies for demand and supply shifters ex ante. Nonetheless, it is instructive to investigate which of the demand and supply shifters that have been suggested in the prior literature are most strongly associated with our estimated demand and supply factors. Thus, in Table 4, we regress 30-year swap spreads ($s_t^{(30)}$), the contribution of the demand factor to spreads ($\hat{A}_d\hat{z}_t^d$), the contribution of the supply factor to spreads ($\hat{A}_w \hat{z}_t^w$), and *PD-UST-Net*$_t$, respectively, on various proxies for demand and supply. Panel A shows the coefficients when these regressions are run in levels, controlling for a linear time trend to partial out any low-frequency trends. Panel B reports the corresponding coefficients when these regressions are run using 13-week changes.

We begin by exploring factors that prior literature suggests may shift end-users' demand for swaps. First, Feldhütter and Lando (2008), Hanson (2014), Malkhozov et al. (2016), and TBAC (2021) argue that the desire of hedged mortgage investors and mortgage servicers to receive the fixed swap rate increases when long-term interest rates fall. A decline in rates increases expected mortgage prepayments, reducing the duration of outstanding mortgages. In response, hedged investors then want to receive the fixed swap rate to add back duration to their portfolios. This view predicts that end-user demand to receive fixed is negatively related to the duration of outstanding MBS. Second, insurers and pension funds receive fixed to manage the gap between the duration of their liabilities and the duration of their on-balance sheet assets (TBAC, 2021). And, Klingler and Sundaresan (2019) argue that pension funds' demand to receive fixed increases as they become more underfunded. Third, the bond issuance of corporates, whom TBAC (2021) also identifies as receivers of fixed, could influence their demand for swaps.

Columns (1), (4), (7), and (10) show the results from regressions of $s_t^{(30)}$, the demand and supply contributions to $s_t^{(30)}$, and *PD-UST-Net*$_t$, respectively, on MBS duration, the pension underfunding measure, and the 3-month moving average of corporate bond issuance.[26] Looking at both levels in Panel A and 3-month changes in Panel B, we find that all three variables are related to our demand factor to some degree, with MBS duration having the most statistically significant relationship with both the 30-year swap spread and its demand-driven component. This result echoes earlier findings of Feldhütter and Lando (2008) and Hanson (2014) that hedging demand from MBS investors is an important driver of the swap spreads.[27] However, the importance of the MBS variable in these regressions does not imply that MBS investors play an exclusive role in the swap market. Domanski et al. (2017) argue that insurers and pension funds also engage in convexity-driven hedging: when interest rates fall, they add duration by receiving the fixed swap rate to offset the increasing mismatch between their assets and liabilities. Thus, the MBS duration variable may also capture demand from insurers and pension funds.

Turning to factors that shift intermediary supply, our theory suggests that intermediaries' supply of swaps should be positively related to the balance-sheet capacity of broker-dealers and hedge funds. In columns (2), (5), (8), and (11), we consider two proxies for the balance-sheet capacity of intermediaries: the broker-dealer leverage factor from Adrian et al. (2014) and the logarithm of the HFRX Global Hedge Fund Index, which follows the approach in Liu (2020). As shown in Panel A of Table 4, both of these variables are positively related to level of swap spreads and the supply-driven component of spreads as one might expect. However, both variables are also positively related to the demand-driven component spreads and the relationships are less clear-cut

[26]The Klingler and Sundaresan (2019) pension underfunding variable and the broker-dealer leverage factor from Adrian et al. (2014) only available at a quarterly frequency. We create weekly versions of these variables by interpolating linearly between the quarter-ending values.

[27]Figure D-2 in the Internet Appendix shows the time series of MBS duration and the demand factor: the series are almost mirror images of each other.

in Panel B when these regressions are run in 3-month changes.

Finally, we relate our estimated end-user demand and intermediary supply factors to general conditions in fixed-income markets. Specifically, columns (3), (6), (9), and (12) in Table 4 report regressions involving the estimated term premium component of 10-year Treasury yields from Adrian et al. (2013), the VXTY option-implied Treasury yield volatility index, the absolute deviation from 3-month covered interest-rate parity averaged across the EUR-USD, GBP-USD, and JPY-USD currency pairs, and the Hu et al. (2013) yield curve noise measure which captures illiquidity in the Treasury market.

A noteworthy pattern that emerges from these regressions is the strong positive relationship between the term premium and the demand-driven component of the swap spread, which reflects the positive relationship between the term premium and swap spreads and the strong negative relationship between the term premium and $\hat{d}_t = \textit{PD-UST-Net}_t$. This pattern is consistent with the role played by MBS investors in the swap market. As discussed above, MBS duration is negatively related to the demand to receive fixed. In turn, Hanson (2014) and Malkhozov et al. (2016) show that MBS duration positively predicts Treasury returns. The close connection between the demand factor, MBS duration, and term premium is illustrated on Figure D-2.[28]

Panel A of Table 4 also shows that there is a strong association between the absolute magnitude of CIP deviations and the level of swap spreads, with swap spreads becoming more negative when CIP deviations are larger in magnitude. This fact is also noted in Siriwardane et al. (2021) and Du et al. (2022). One natural interpretation is that short-dated CIP deviations reflect the shadow cost of intermediary capital. Interestingly, our analysis suggests that this low-frequency relationship between the level of swap spreads and CIP deviations works through both the supply- and the demand-driven components of spreads. However, Panel B shows that this relationship largely disappears when we look at 3-month changes. As argued by Siriwardane et al. (2021), if these two markets are partially segmented, the shadow cost of intermediary capital reflected in CIP deviations need not be the exact same as the shadow cost reflected in swap spreads. Thus, to the extent that these markets are better integrated at low-frequencies than at high-frequencies—e.g., due to slow-moving capital effects, this may explain why the relationship is much stronger at low frequencies.

Impulse-response functions and estimated half-lives. Panel A of Figure 4 shows the impulse response functions from (37). These IRFs enable us to estimate the half-lives of the responses of the latent demand and supply factors (i.e., the elements of $\mathbf{z}_t$) to structural shocks to demand or supply (i.e., the elements of $\boldsymbol{\varepsilon}_t$). Panel B of 4 shows the corresponding IRFs from (38), allowing us to estimate the half-lives of the responses of observables (i.e., the elements of $\mathbf{y}_t$) to these same structural shocks. In both cases, we define the half-life as the smallest weekly lag length such that the magnitude of the response is less than 50% of its initial value on impact. Note, however, these two sets of half-lives will not be the same since we allow for cross-lag effects between our

[28]While we offer a different demand-and-supply based interpretation, this observation is related to Du et al. (2022) who note that $\textit{PD-UST-Net}_t$ is low when the yield curve is steep. By contrast, the model in Jermann (2020) makes the opposite prediction of what we find in the data. In his model, a higher term premium should be associated with higher values of $\textit{PD-UST-Net}_t$ and lower (more negative) values of $s_t^{(30)}$.

structural factors—i.e., since the $\mathbf{D}_l$ are not diagonal, shocks to supply may affect future demand and vice versa.

We begin by considering shocks to intermediary supply (ε_t^w). We find that the response of the latent supply factor (z_t^w) to its own impulse (ε_t^w) subsides by half after 26 weeks. This estimated half-life can be compared to Goldberg and Nozawa (2021) and Fontaine et al. (2023), who consider shocks to financial intermediaries with a broadly similar interpretation to our supply shocks. Goldberg and Nozawa (2021) find that shocks to intermediary liquidity supply in the corporate bond market have a half-life of 29 weeks. Similarly, Fontaine et al. (2023) find that the half-life of the response of funding liquidity spreads to shocks to intermediary supply is two quarters—i.e., 26 weeks. Turning to observables, the estimated response of 30-year swap spreads ($s_t^{(30)}$) to an impulse to intermediary supply (ε_t^w) falls by half after 72 weeks. By contrast, the half-life of the response of $PD\text{-}Net\text{-}UST_t$ to shocks to ε_t^w is just 11 weeks. The more persistent effect of supply shocks on swap spreads and the less persistent effect on $PD\text{-}Net\text{-}UST_t$ (as compared to the intermediary supply factor) arises due to an indirect, cross-lag effect. Specifically, the off-diagonals of the $\mathbf{D}_l$ imply that a positive shock to intermediary supply eventually results in an small but persistent decline in end-user demand. This cross-lag effect pushes up swap spreads and reduces intermediary positions, raising the half-life of the former and reducing the half-life of the latter.

How persistent are the responses to shocks to end-user demand (ε_t^d)? Our estimates imply that the response of the latent demand factor (z_t^d) to a demand impulse (ε_t^d) subsides by half after 26 weeks. And, accounting for cross-lag effects, the estimated response of 30-year swap spreads ($s_t^{(30)}$) to a demand impulse subsides by half after 22 weeks. The half-life of the response of $PD\text{-}Net\text{-}UST_t$ to demand shocks is similar at 31 weeks. The nature of the relevant demand shocks is fairly specific to the swap market so, unlike supply shocks, their persistence need not match estimates from other markets. However, we can provide external validation by noting that MBS duration, which we have shown is tightly linked to our estimated demand factor, has a similar persistence. Specifically, using a simple AR(1) model, we estimate that MBS duration has a weekly autocorrelation of 0.975. This implies that the half-life of MBS duration shocks is 27 weeks, which almost perfectly matches the estimated half-life of our demand factor to its own impulse.

Maturity-specific demand factors. In Internet Appendix B.5, we extend our theory to allow end-users' demand to receive fixed to be driven by two separate factors: a factor controlling demand for shorter-dated swaps and a factor controlling demand for longer-dated swaps. As before, we assume that a single group of risk-averse and leverage-constrained swap intermediaries absorbs the net end user demand at each maturity, hedging the resulting interest-rate risk by taking offsetting positions in like-maturity Treasuries. The spreads on n-year swap spreads take the form

$$s_t^{(n)} = A_0^{(n)} + A_m^{(n)} z_t^m + A_{d,S}^{(n)} z_t^{d,S} + A_{d,L}^{(n)} z_t^{d,L} + A_w^{(n)} z_t^w, \tag{41}$$

where $z_t^{d,S}$ and $z_t^{d,L}$ are the short- and long-dated demand factors.

As explained in Internet Appendix C.2, we set identify the short-dated demand factor, the long-dated demand factor, and supply shocks in a structural VAR that includes the 5-year swap spread ($s_t^{(5)}$), the 30-year swap spread ($s_t^{(30)}$), primary dealers' net position in Treasuries with maturities of 6 years or less ($\hat{d}_t^S = -\hat{x}_t^S$), and primary dealers' net position in Treasuries with maturities over 6 years ($\hat{d}_t^L = -\hat{x}_t^L$). We achieve set identification by imposing a set of sign and monotonicity restrictions implied by our theory. Specifically, since our theory does not make unambiguous predictions about how $z_t^{d,S}$ and $z_t^{d,L}$ should impact the shape of the swap spread curve, our key identifying assumption is simply that shocks to short-dated (long-dated) end-user demand has a greater impact on primary dealers' net position in short-dated (long-dated) Treasuries.

Our estimated short-dated demand factor explains 39% of the residual variation in 5-year swap spreads and only 6% of the variation in the 30-year swap spread at a 1-year horizon. By contrast, the long-dated demand factor explains only 3% of the 1-year residual variation in 5-year spreads and 70% of the variation in 30-year spreads.

To shed light on the underlying drivers of end-user demand for short- and long-dated swaps, Table D-3 of the Internet Appendix reports regressions of 3-month changes in our two estimated demand factors on changes in MBS duration, the Klingler and Sundaresan (2019) pension underfunding measure, and corporate bond issuance. Our results suggest that MBS duration and other related convexity-driven flows play a major role in driving the demand for both short- and long-maturity swaps. However, consistent with the argument in Klingler and Sundaresan (2019), we find that pension underfunding primarily influences the demand for long-dated swaps.

5 Conclusion

We develop a tractable model in which long-term swap spreads are shaped by end users' net demand for and constrained intermediaries supply of swaps. Even if swaps are completely redundant, non-zero spreads arise when constrained intermediaries must accommodate end-user demand for swaps in equilibrium. Long-term swap spreads reflect both compensation for using scarce intermediary capital and compensation for bearing convergence risk—i.e., the risk that spreads will move against intermediaries due to an unexpected future imbalance between demand and supply.

We find that a proxy for intermediated quantities in the swap market—primary dealers' net position in U.S. Treasuries—flipped sign at the height of the GFC, at the same time when long-term swap spreads turned negative. Consistent with the importance of convergence risk, this proxy for intermediated quantities predicts the excess returns on long-dated swap spread trades. Using this quantity proxy, we identify shifts in end-user demand and intermediary supply, and find that both contribute significantly to the volatility of long-term swap spreads, while only demand is powerful in forecasting short-term returns on swap spread arbitrage.

While we have applied our framework to study long-dated swap spreads, in the future it could be used to explore other long-dated near-arbitrage spreads, including deviations from covered interest rate parity and the CDS-bond basis.

References

Adrian, T., R. K. Crump, and E. Moench (2013). Pricing the term structure with linear regressions. *Journal of Financial Economics 110*(1), 110–138.

Adrian, T., E. Etula, and T. Muir (2014). Financial Intermediaries and the Cross-Section of Asset Returns. *The Journal of Finance 69*(6), 2557–2596.

Anderson, C. and W. Liu (2021). Inferring intermediary risk exposure from trade. Working paper, Federal Reserve Board.

Arias, J. E., J. F. Rubio-Ramirez, and D. F. Waggoner (2018). Inference based on structural vector autoregressions identified with sign and zero restrictions: Theory and applications. *Econometrica 86*(2), 685–720.

Augustin, P., M. Chernov, L. Schmid, and D. Song (2021). Benchmark interest rates when the government is risky. *Journal of Financial Economics 140*(1), 74–100.

Begenau, J., M. Piazzesi, and M. Schneider (2020). Banks' risk exposures. Working paper, Stanford University.

Boyarchenko, N., T. M. Eisenbach, P. Gupta, O. Shachar, and P. Van Tassel (2020). Bank-intermediated arbitrage. Staff Report 858, Federal Reserve Bank of New York.

Boyarchenko, N., P. Gupta, N. Steele, and J. Yen (2018). Negative swap spreads. *Federal Reserve Bank of New York Economic Policy Review 24*(2), 1–14.

Brunnermeier, M. K. and L. H. Pedersen (2009). Market liquidity and funding liquidity. *Review of Financial Studies 22*(6), 2201–2238.

Campbell, J. Y. (2018). *Financial Decisions and Markets: A Course in Asset Pricing*. Princeton University Press.

Campbell, J. Y. and R. J. Shiller (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies 1*(3), 195–228.

Campbell, J. Y. and R. J. Shiller (1991). Yield spreads and interest rate movements: A bird's eye view. *Review of Economic Studies 58*(3), 495–514.

Chen, H., S. Joslin, and S. X. Ni (2018). Demand for crash insurance, intermediary constraints, and risk premia in financial markets. *Review of Financial Studies 32*(1), 228–265.

Cieslak, A. and H. Pang (2021). Common shocks in stocks and bonds. *Journal of Financial Economics 142*(2), 880–904.

Cohen, L., K. B. Diether, and C. J. Malloy (2007). Supply and demand shifts in the shorting market. *Journal of Finance 62*(5), 2061–2096.

Collin-Dufresne, P. and B. Solnik (2001). On the term structure of default premia in the swap and libor markets. *Journal of Finance 56*, 1095–1115.

Cortes, F. (2003). Understanding and modelling swap spreads. Quarterly bulletin, Bank of England.

De Long, B. J., A. Shleifer, L. H. Summers, and R. J. Waldmann (1990). Noise trader risk in financial markets. *Journal of Political Economy 98*(4), 703–738.

Domanski, D., H. S. Shin, and V. Sushko (2017). The hunt for duration: Not waving but drowning? *IMF Economic Review 65*(1), 113–153.

Drechsler, I., A. Savov, and P. Schnabl (2021). Banking on deposits: Maturity transformation without interest rate risk. *Journal of Finance 76*(3), 1091–1143.

Driscoll, J. C. and R. A. Judson (2013). Sticky deposit rates. Working paper, Board of Governors of the Federal Reserve System.

Du, W., B. Hébert, and A. W. Huber (2022). Are intermediary constraints priced? *Review of Financial Studies*, forthcoming.

Du, W., B. Hébert, and W. Li (2022). Intermediary balance sheet and the treasury yield curve. Working paper, University of Chicago.

Du, W., A. Tepper, and A. Verdelhan (2018). Deviations from Covered Interest Rate Parity. *Journal of Finance 73*(3), 915–957.

Duarte, J., F. A. Longstaff, and F. Yu (2006). Risk and return in fixed-income arbitrage: Nickels in front of a steamroller? *Review of Financial Studies 20*(3), 769–811.

Duffie, D. (2017). Post-crisis bank regulations and financial market liquidity. Technical report, Banca d'Italia.

Duffie, D. and K. J. Singleton (1997). An econometric model of the term structure of interest-rate swap yields. *Journal of Finance 52*, 1287–1321.

Fama, E. and R. R. Bliss (1987). The information in long-maturity forward rates. *Journal of Financial Economics 77*(4), 680–692.

Feldhütter, P. and D. Lando (2008). Decomposing swap spreads. *Journal of Financial Economics 88*, 375–405.

Fleming, M. J. and J. V. Rosenberg (2008). How do treasury dealers manage their positions? Technical report.

Fontaine, J.-S., R. Garcia, and S. Gungor (2023). Intermediary leverage shocks and funding conditions. Working paper, Bank of Canada.

Fry, R. and A. Pagan (2005). Some issues in using vars for macroeconometric research. Technical report.

Fry, R. and A. Pagan (2011). Sign restrictions in structural vector autoregressions: A critical review. *Journal of Economic Literature 49*(4), 938–960.

Gârleanu, N. and L. H. Pedersen (2011). Margin-based asset pricing and deviations from the law of one price. *Review of Financial Studies 24*, 1980–2022.

Goldberg, J. and Y. Nozawa (2021). Liquidity supply in the corporate bond market. *Journal of Finance 76*(2), 755–796.

Gourinchas, P.-O., W. Ray, and D. Vayanos (2020). A preferred-habitat model of term premia and currency risk. Working paper, London School of Economics.

Greenwood, R., S. G. Hanson, J. C. Stein, and A. Sunderam (2017). Strengthening and streamlining bank capital regulation. *Brookings Papers on Economic Activity 48*(2), 479–565.

Greenwood, R., S. G. Hanson, J. C. Stein, and A. Sunderam (2022). A quantity-driven theory of term premiums and exchange rates. Working paper, Harvard Business School.

Greenwood, R. and D. Vayanos (2014). Bond supply and excess bond returns. *Review of Financial Studies 27*(3), 663–713.

Gromb, D. and D. Vayanos (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics 66*(2), 361–407.

Haddad, V. and D. Sraer (2020). The banking view of bond risk premia. *Journal of Finance 75*(5), 2465–2502.

Hanson, S. G. (2014). Mortgage convexity. *Journal of Financial Economics 113*(2), 270–299.

He, Z., B. Kelly, and A. Manela (2017). Intermediary Asset Pricing: New Evidence from Many Asset Classes. *Journal of Financial Economics 126*(1), 1–35.

He, Z. and A. Krishnamurthy (2013). Intermediary Asset Pricing. *American Economic Review 103*(2), 732–770.

He, Z., S. Nagel, and Z. Song (2022). Treasury inconvenience yields during the covid-19 crisis. *Journal of Financial Economics 143*(1), 57–79.

Hu, G. X., J. Pan, and J. Wang (2013). Noise as information for illiquidity. *Journal of Finance 68*, 2341–2382.

Jermann, U. (2020). Negative swap spreads and limited arbitrage. *Review of Financial Studies 33*(1), 212–238.

Kekre, R., M. Lenel, and F. Maindardi (2022). Monetary policy, segmentation, and the term structure. Working paper, University of Chicago.

Kiefer, N. M. and T. J. Vogelsang (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory 21*, 1130–1164.

Klingler, S. and S. Sundaresan (2019). An explanation of negative swap spreads: Demand for duration from underfunded pension plans. *Journal of Finance 74*(2), 675–710.

Kondor, P. (2009). Risk in dynamic arbitrage: the price effects of convergence trading. *Journal of Finance 64*, 631–655.

Krishnamurthy, A. and A. Vissing-Jorgensen (2012). The aggregate demand for treasury debt. *Journal of Political Economy 120*(2), 233–267.

Lang, L. H., R. H. Litzenberger, and A. L. Liu (1998). Determinants of interest rate swap spreads. *Journal of Banking and Finance 22*, 1507–1532.

Liu, J. (2020). Comovement in arbitrage limits. Working paper, University of Hong Kong.

Liu, J., F. A. Longstaff, and R. E. Mandell (2006). The market price of risk in interest rate swaps: The roles of default and liquidity risks. *Journal of Business 79*, 2337–2359.

Lowenstein, R. (2000). *When Genius Failed: The Rise and Fall of Long-Term Capital Management*. New York: Random House.

Malkhozov, A., P. Mueller, A. Vedolin, and G. Venter (2016). Mortgage risk and the yield curve. *Review of Financial Studies 29*(5), 1220–1253.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*, 703–708.

O'Brien, J. and J. Berkowitz (2007). Estimating bank trading risk: A factor model approach. In M. Carey and R. M. Stulz (Eds.), *The Risks of Financial Institutions*. University of Chicago Press.

Perli, R. and B. P. Sack (2003). Does mortgage hedging amplify movements in long-term interest rates? *Journal of Fixed Income 13*(3), 7 – 17.

Siriwardane, E., A. Sunderam, and J. Wallen (2021). Segmented arbitrage. Working paper, Harvard Business School.

TBAC (2021, February). Treasury borrowing advisory committee report to treasury. Technical report.

Uhlig, H. (2005). What are the effects of monetary policy on output? results from an agnostic identification procedure. *Journal of Monetary Economics 52*, 381–419.

Vayanos, D. and J.-L. Vila (2021). A preferred-habitat model of the term structure of interest rates. *Econometrica 89*, 77–112.

Table 1: **Forecasting returns to the 30-year swap spread trade from 2009m1 to 2020m6:** This table reports slope coefficients from regressions of the form

$$r^{s(30)}_{t\to t+h} = \alpha + \beta_1 \cdot \textit{PD-UST-Net}_t + \beta_2 \cdot \textit{PD-UST-Credit}_t + \gamma_1 \cdot \textit{ST-OIS-Spread}_t + \boldsymbol{\gamma}_2'\mathbf{x}_t + \epsilon_{t\to t+h}.$$

We regress the returns to the 30-year swap spread trade on Primary Dealers' net position in nominal coupon-bearing Treasury securities (*PD-UST-Net*$_t$), Primary Dealers' outstanding balances of Treasury securities in through financing arrangements (*PD-UST-Credit*$_t$), the spread between the 3-month OIS rate and the 3-month T-bill rate (*ST-OIS-Spread*$_t$), and controls. Controls include the Federal Reserve's holdings of coupon-bearing Treasury securities, the 12-month moving average of the gross issuance of coupon-bearing Treasuries, and Primary Dealers' net position in corporate bonds. The holding period is equal to $h = 13$ weeks (3 months) in columns 1 to 4 and to $h = 52$ weeks (12 months) in columns 5 to 8. Data are weekly and run from January 2009 to June 2020 in Panel A and from January 2009 to June 2018 in Panel B. Newey and West (1987) t-statistics computed with 20 lags for 3-month return regressions and 78 lags for 12-month return regressions are reported in parentheses. We compute the associated p-values using the fixed-b asymptotic theory of Kiefer and Vogelsang (2005). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	h = 3-month returns				h = 12-month returns			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Sample from 2009m1 to 2020m6								
PD-UST-Net$_t$	0.172	-1.064	0.243	-0.837	-1.747	-5.039***	-1.226	-4.053**
	(0.47)	(1.47)	(0.62)	(1.12)	(1.57)	(3.33)	(1.04)	(2.81)
PD-UST-Credit$_t$			-0.169	-0.392***			-1.243**	-1.703***
			(1.09)	(2.78)			(3.06)	(5.00)
ST-OIS-Spread$_t$	0.833	0.362	0.163	0.853	-17.155	-26.064**	-22.096	-23.931**
	(0.22)	(0.09)	(0.04)	(0.23)	(1.18)	(2.45)	(1.70)	(2.61)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Adjusted R^2	-0.001	0.043	0.010	0.086	0.075	0.141	0.231	0.327
N	599	599	599	599	599	599	599	599
Panel B: Sample from 2009m1 to 2018m6								
PD-UST-Net$_t$	-1.580***	-3.036***	-1.610***	-2.901***	-4.965**	-7.437***	-5.142***	-6.752***
	(3.75)	(4.53)	(3.79)	(4.39)	(2.72)	(4.19)	(3.45)	(5.72)
PD-UST-Credit$_t$			-0.200	-0.294**			-1.179*	-1.491***
			(1.11)	(2.41)			(2.18)	(5.49)
ST-OIS-Spread$_t$	8.064**	4.933	7.438**	4.473	-1.440	-15.183	-5.133	-17.518
	(2.20)	(1.57)	(2.14)	(1.47)	(0.07)	(1.14)	(0.31)	(1.60)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Adjusted R^2	0.132	0.178	0.149	0.205	0.216	0.379	0.348	0.526
N	495	495	495	495	495	495	495	495

Table 2: **Forecasting returns to 30-year and 10-year swap spread trades from 2009m1 to 2018m6:** This table reports the slope coefficients from regressions of the form

$$r^{s(n)}_{t\to t+h} = \alpha + \beta_1 \cdot \textit{PD-UST-Net-Alt}_t + \beta_2 \cdot \textit{PD-UST-Credit}_t + \gamma_1 \cdot \textit{ST-OIS-Spread}_t + \boldsymbol{\gamma}_2' \mathbf{x}_t + \epsilon_{t\to t+h}$$

for $n = 10$ and 30. We regress returns to swap spread trades on alternative measures of Primary Dealers' net position in Treasury securities (*PD-UST-Net-Alt*$_t$), Primary Dealers' outstanding balances of Treasury securities in through financing arrangements (*PD-UST-Credit*$_t$), and the spread between the 3-month OIS rate and the 3-month T-bill rate (*ST-OIS-Spread*$_t$). We show results for our baseline net exposure measure (*PD-UST-Net*$_t$), a measure focusing on Treasuries maturing in 11 or more years (*PD-UST*$^{11+}$*-Net*$_t$), and a measure of levered exposure (*PD-UST-Lev*$_t$). Returns are for the $n = 30$-year swap spread trade in Panel A and for the $n = 10$-year swap spread trade in Panel B, and the holding period is equal to $h = 13$ weeks (3 months) in columns 1 to 4 and to $h = 52$ weeks (12 months) in columns 5 to 8. Data are weekly and run from January 2009 to June 2018. Newey and West (1987) t-statistics computed with 20 lags for 3-month return regressions and 78 lags for 12-month return regressions are reported in parentheses. We compute the associated p-values using the fixed-b asymptotic theory of Kiefer and Vogelsang (2005). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	$h = 3$-month returns			$h = 12$-month returns		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Returns on the 30-year swap spread trade, $r^{s(30)}_{t\to t+h}$						
PD-UST-Net$_t$	-1.610***			-5.142***		
	(3.79)			(3.45)		
PD-UST^{11+}*-Net*$_t$		-7.930***			-16.262*	
		(4.05)			(2.24)	
PD-UST-Lev$_t$			-1.145***			-2.993**
			(3.58)			(2.53)
PD-UST-Credit$_t$	-0.200	-0.263	-0.298	-1.179*	-1.283**	-1.424*
	(1.11)	(1.50)	(1.67)	(2.18)	(2.46)	(2.36)
ST-OIS-Spread$_t$	7.438**	8.482**	7.050*	-5.133	-5.058	-7.194
	(2.14)	(2.33)	(1.96)	(0.31)	(0.30)	(0.43)
Adjusted R^2	0.149	0.164	0.133	0.348	0.228	0.252
N	495	495	495	495	495	495
Panel B: Returns on the 10-year swap spread trade, $r^{s(10)}_{t\to t+h}$						
PD-UST-Net$_t$	-0.227			-1.017*		
	(1.38)			(2.34)		
PD-UST^{11+}*-Net*$_t$		-1.890***			-2.523	
		(3.71)			(1.36)	
PD-UST-Lev$_t$			-0.170			-0.621
			(1.51)			(1.68)
PD-UST-Credit$_t$	-0.134**	-0.151***	-0.149***	-0.348**	-0.361**	-0.400**
	(2.73)	(3.13)	(2.94)	(3.25)	(2.68)	(3.15)
ST-OIS-Spread$_t$	1.385	1.810	1.340	-2.103	-2.337	-2.474
	(1.18)	(1.68)	(1.15)	(0.65)	(0.70)	(0.74)
Adjusted R^2	0.097	0.138	0.096	0.267	0.182	0.224
N	495	495	495	495	495	495

Table 3: **Fama-Bliss style forecasting regressions:** This table reports the slope coefficients from regressions of the form

$$r^{s(n)}_{t\to t+h} - r^{s(n')}_{t\to t+h} = \alpha + \beta \cdot (s^{(n)}_t - s^{(n')}_t) + \epsilon_{t\to t+h},$$

where $n > n'$. In words, we regress the excess returns on a longer-dated swap spread trade over a shorter-dated trade ($r^{s(n)} - r^{s(n')}$) on the corresponding difference between longer- and shorter-dated swap spreads ($s^{(n)}_t - s^{(n')}_t$). The holding period is equal to $h = 13$ weeks (3 months) in columns 1 to 3 and to $h = 52$ weeks (12 months) in columns 4 to 6. Data are weekly and run from January 2009 to June 2020 in Panel A and from January 2009 to June 2018 in Panel B. Newey and West (1987) t-statistics computed with 20 lags for 3-month return regressions and 78 lags for 12-month return regressions are reported in parentheses. We compute the associated p-values using the fixed-b asymptotic theory of Kiefer and Vogelsang (2005). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	$h = 3$-month returns			$h = 12$-month returns		
	(1) $r^{s(30)} - r^{s(1)}$	(2) $r^{s(10)} - r^{s(1)}$	(3) $r^{s(30)} - r^{s(10)}$	(4) $r^{s(30)} - r^{s(1)}$	(5) $r^{s(10)} - r^{s(1)}$	(6) $r^{s(30)} - r^{s(10)}$
Panel A: Sample from 2009m1 to 2020m6						
$s^{(30)}_t - s^{(1)}_t$	2.550***			12.577***		
	(3.03)			(4.97)		
$s^{(10)}_t - s^{(1)}_t$		1.343***			2.988**	
		(3.58)			(2.51)	
$s^{(30)}_t - s^{(10)}_t$			5.674***			21.333***
			(4.05)			(6.58)
Adjusted R^2	0.077	0.101	0.140	0.422	0.163	0.454
N	599	599	599	599	599	599
Panel B: Sample from 2009m1 to 2018m6						
$s^{(30)}_t - s^{(1)}_t$	2.061**			11.143***		
	(2.33)			(4.36)		
$s^{(10)}_t - s^{(1)}_t$		1.284***			2.940*	
		(3.19)			(2.30)	
$s^{(30)}_t - s^{(10)}_t$			4.726***			18.623***
			(3.12)			(5.62)
Adjusted R^2	0.063	0.098	0.120	0.381	0.168	0.384
N	495	495	495	495	495	495

Table 4: **Demand and supply drivers of 30-year swap spreads from 2009m1 to 2018m6:** This table reports the slope coefficients from regressions of, respectively, the 30-year swap spread (columns 1 to 3), the demand contribution to spreads (columns 4 to 6), the supply contribution to spreads (columns 7 to 9), and $PD\text{-}UST\text{-}Net_t$ (columns 10 to 12) on the modified duration of the Barclays U.S. MBS index ($MBS\text{-}Duration_t$), the Klingler and Sundaresan (2019) pension underfunding factor ($Pension\text{-}UFR_t$), corporate bond issuance ($Corp\text{-}Issuance_t$), the broker-dealer leverage factor from Adrian et al. (2014) ($BD\text{-}Lev_t$), the logarithm of the HFRX Global Hedge Fund Index ($\ln(HFR_t)$), the Adrian et al. (2013) term premium ($Term\text{-}Prem\text{-}ACM_t$), the VXTY Treasury volatility index ($UST\text{-}Implied\text{-}Vol_t$), the absolute deviation from 3-month covered interest-rate parity (computed as in Du et al. (2018)) averaged across the EUR-USD, GBP-USD, and JPY-USD currency pairs ($Avg\text{-}Cip\text{-}Basis_t$), and the Hu et al. (2013) yield curve fitting error ($UST\text{-}Curve\text{-}Noise_t$). Variables are in levels in Panel A and we control for a time trend (t) and in 13-week changes in Panel B. Data are weekly and run from January 2009 to June 2018. Newey and West (1987) t-statistics are reported in parentheses and are computed with 78 lags in Panel A and with 20 lags in Panel B. We compute the associated p-values using the fixed-b asymptotic theory of Kiefer and Vogelsang (2005). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	$s_t^{(30)}$			$s_t^{(30)}$: Demand contrib.			$s_t^{(30)}$: Supply contrib.			$PD\text{-}UST\text{-}Net_t$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Regressions in Levels												
$MBS\text{-}Duration_t$	8.984***			6.297***			2.687			-11.523*		
	(4.65)			(7.49)			(1.66)			(2.16)		
$Pension\text{-}UFR_t$	-1.448			-0.592			-0.857			-0.795		
	(1.42)			(1.16)			(1.23)			(0.38)		
$Corp\text{-}Issuance_t$	-0.007			-0.041			0.034			0.235		
	(0.08)			(0.49)			(0.62)			(0.65)		
$\ln(BD\text{-}Lev_t)$		147.138***			43.275			103.863**			186.131	
		(3.21)			(1.69)			(3.24)			(1.67)	
$\ln(HFR_t)$		2.777***			1.894***			0.883*			-3.232**	
		(3.67)			(4.41)			(2.22)			(3.08)	
$Term\text{-}Prem\text{-}ACM_t$			2.866			5.729*			-2.863			-26.972*
			(0.78)			(2.18)			(1.05)			(2.20)
$UST\text{-}Implied\text{-}Vol_t$			-0.052			0.439			-0.491			-2.909
			(0.05)			(0.53)			(0.71)			(0.84)
$Avg\text{-}CIP\text{-}Basis_t$			-0.631***			-0.330**			-0.301***			0.107
			(3.53)			(2.60)			(3.35)			(0.27)
$UST\text{-}Curve\text{-}Noise_t$			-0.232			-0.131			-0.102			0.097
			(0.43)			(0.39)			(0.29)			(0.07)
Adjusted R^2	0.479	0.459	0.440	0.606	0.606	0.563	0.286	0.368	0.349	0.507	0.619	0.561
N	495	495	495	495	495	495	495	495	495	495	495	495
Panel B: Regressions in 3-month changes												
$\Delta_{13}MBS\text{-}Duration_t$	3.728***			3.062***			0.666			-7.594**		
	(5.01)			(4.93)			(1.08)			(2.44)		
$\Delta_{13}Pension\text{-}UFR_t$	-1.056			-1.415***			0.359			5.578***		
	(1.46)			(2.61)			(1.00)			(3.14)		
$\Delta_{13}Corp\text{-}Issuance_t$	-0.114**			-0.088**			-0.025			0.200		
	(2.36)			(2.14)			(0.90)			(1.23)		
$\Delta_{13}\ln(BD\text{-}Lev_t)$		-0.219			-13.906			13.687			86.395	
		(0.01)			(0.45)			(0.61)			(0.71)	
$\Delta_{13}\ln(HFR_t)$		1.097**			1.364***			-0.267			-5.132***	
		(2.24)			(4.39)			(1.04)			(5.57)	
$\Delta_{13}Term\text{-}Prem\text{-}ACM_t$			3.087			5.686***			-2.599			-26.014***
			(1.06)			(2.95)			(1.41)			(3.47)
$\Delta_{13}UST\text{-}Implied\text{-}Vol_t$			1.030*			0.229			0.801*			1.766
			(1.79)			(0.42)			(2.10)			(0.76)
$\Delta_{13}Avg\text{-}CIP\text{-}Basis_t$			-0.041			-0.037			-0.004			0.103
			(0.46)			(0.56)			(0.09)			(0.45)
$\Delta_{13}UST\text{-}Curve\text{-}Noise_t$			0.007			0.247			-0.240			-1.523
			(0.01)			(0.40)			(0.96)			(0.95)
Adjusted R^2	0.196	0.080	0.042	0.290	0.233	0.114	0.022	0.018	0.040	0.179	0.226	0.133
N	482	482	482	482	482	482	482	482	482	482	482	482

Table 5: **Forecasting swap spread trade returns using the demand and supply factors:** This table reports the slope coefficients from regressions of the form

$$r^{s(n)}_{t\to t+h} = \alpha + \beta_1 \cdot \mathit{Demand}_t + \beta_2 \cdot \mathit{Supply}_t + \epsilon_{t\to t+h}$$

for $n = 10$ and 30 years. In words, we regress the returns to the n-year swap spread trade on the estimated demand factor (Demand_t) and the supply factor (Supply_t). We also report the coefficients from the corresponding regressions on $\mathit{PD\text{-}UST\text{-}Net}_t$ and $s^{(30)}_t$ which have the same R^2 by construction. Returns are for the 30-year swap spread trade in Panel A and for the 10-year swap spread trade in Panel B. The holding period is equal to $h = 13$ weeks (3 months) in columns 1 and 2 and to $h = 52$ weeks (12 months) in columns 3 and 4. Data are weekly and run from January 2009 to June 2018. Newey and West (1987) t-statistics computed with 20 lags for 3-month return regressions and 78 lags for 12-month return regressions are reported in parentheses. We compute the associated p-values using the fixed-b asymptotic theory of Kiefer and Vogelsang (2005). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	h = 3-month returns		h = 12-month returns	
	(1)	(2)	(3)	(4)
Panel A: Returns on the 30-year swap spread trade, $r^{s(30)}_{t\to t+h}$				
Demand_t	-13.752***		-55.944***	
	(3.51)		(6.06)	
Supply_t	-1.646		9.666	
	(0.37)		(0.71)	
$\mathit{PD\text{-}UST\text{-}Net}_t$		-1.096***		-3.304**
		(2.75)		(2.86)
$s^{(30)}_t$		2.677*		14.531***
		(1.92)		(3.93)
Adjusted R^2	0.130	0.130	0.508	0.508
N	495	495	495	495
Panel B: Returns on the 10-year swap spread trade, $r^{s(10)}_{t\to t+h}$				
Demand_t	-2.684*		-11.707***	
	(1.93)		(4.41)	
Supply_t	2.063		3.356	
	(1.03)		(0.85)	
$\mathit{PD\text{-}UST\text{-}Net}_t$		-0.046		-0.597
		(0.23)		(1.44)
$s^{(30)}_t$		1.053**		3.337***
		(2.46)		(4.56)
Adjusted R^2	0.069	0.069	0.317	0.317
N	495	495	495	495

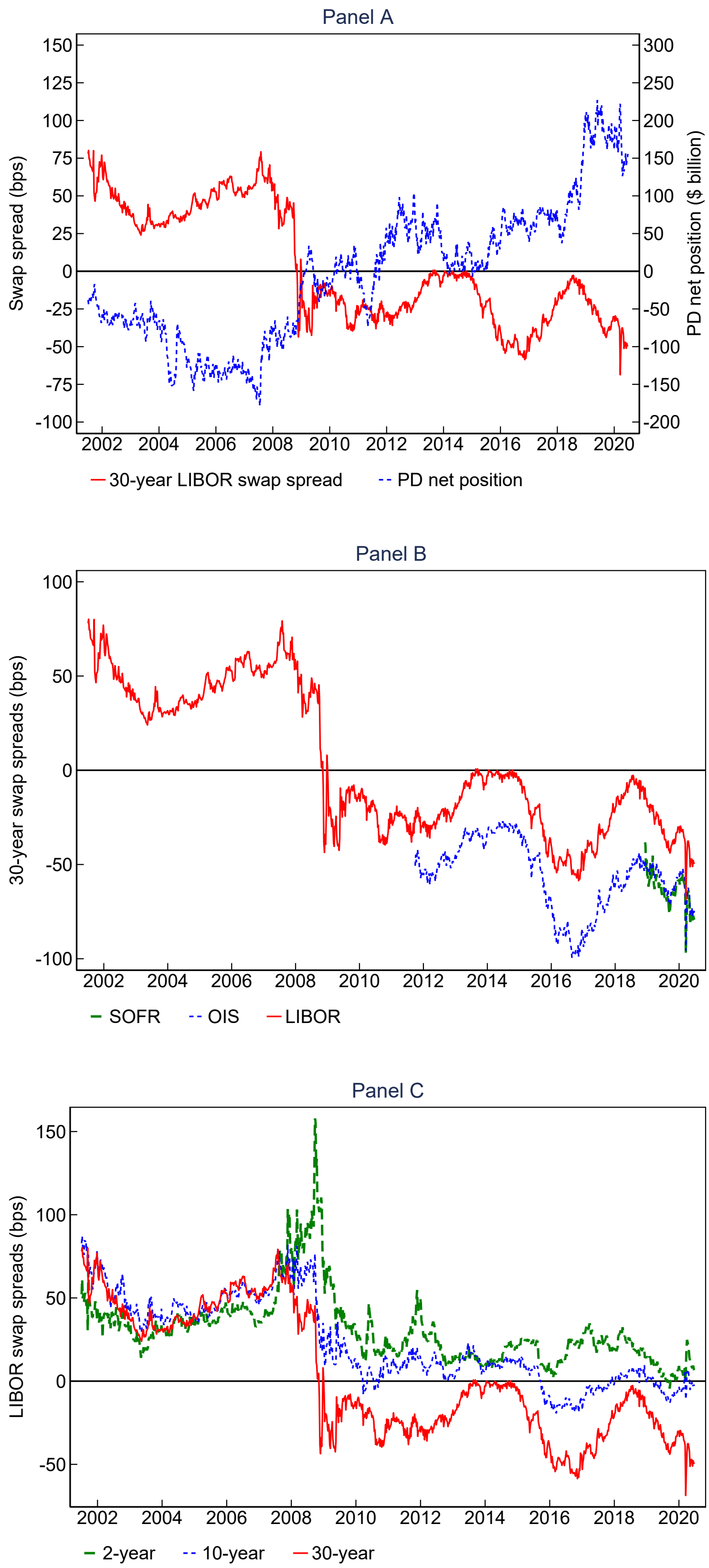


Figure 1: **Swap spreads and primary dealer's net Treasury position:** Panel A shows the 30-year LIBOR swap spread and primary dealers' net position in coupon-bearing Treasury securities ($PD\text{-}UST\text{-}Net_t$). Panel B shows 30-year SOFR, 30-year OIS, and 30-year LIBOR swap spreads. Panel C shows 2-year LIBOR, 10-year LIBOR, and 30-year LIBOR swap spreads. Data are weekly and run from July 2001 to June 2020.

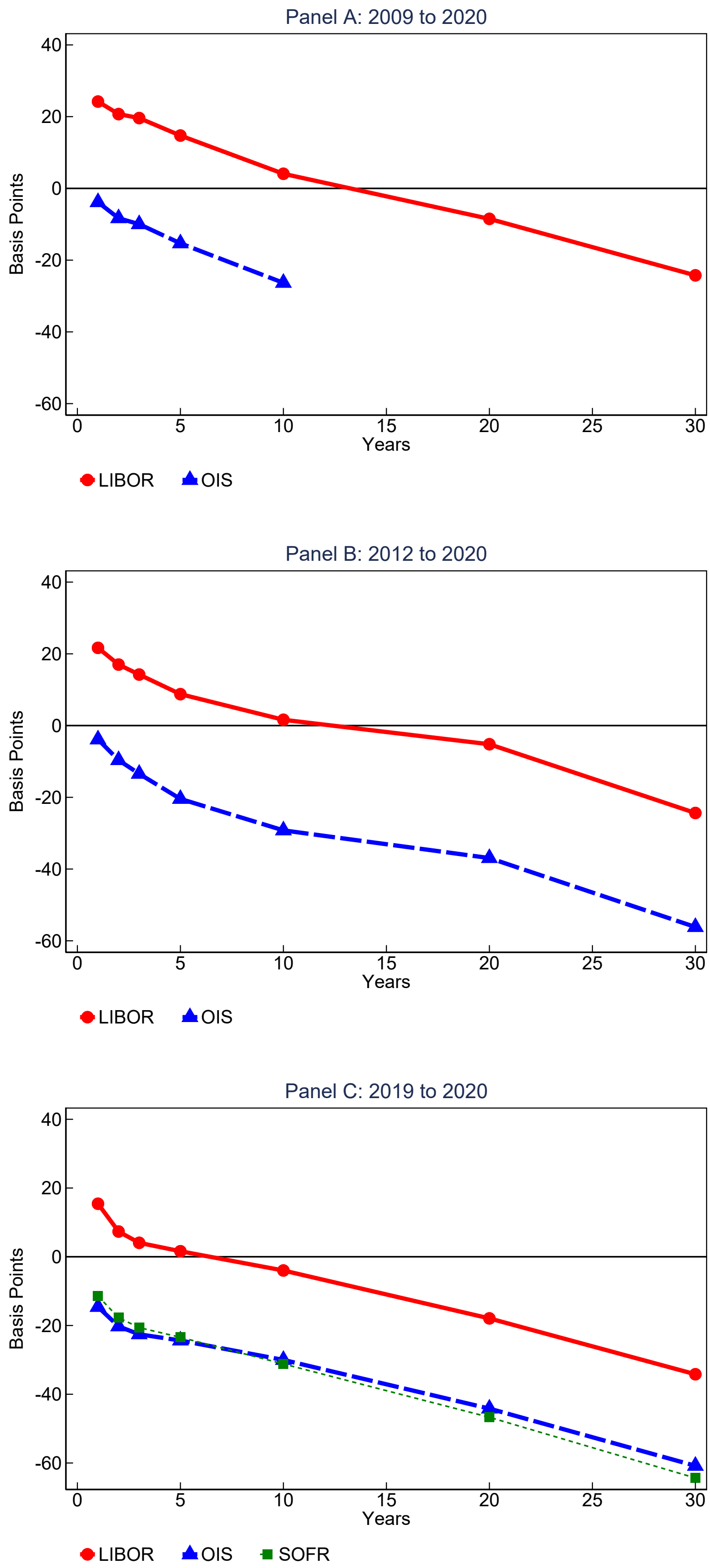


Figure 2: **Term structures of swap spreads:** This figure shows average LIBOR, OIS, and SOFR swap spreads at 1-, 2-, 3-, 5-, 10-, 20 and 30-year maturities for subsamples starting in January 2009 (Panel A), January 2013 (Panel B) and January 2019 (Panel C), and ending in June 2020. Data are weekly.

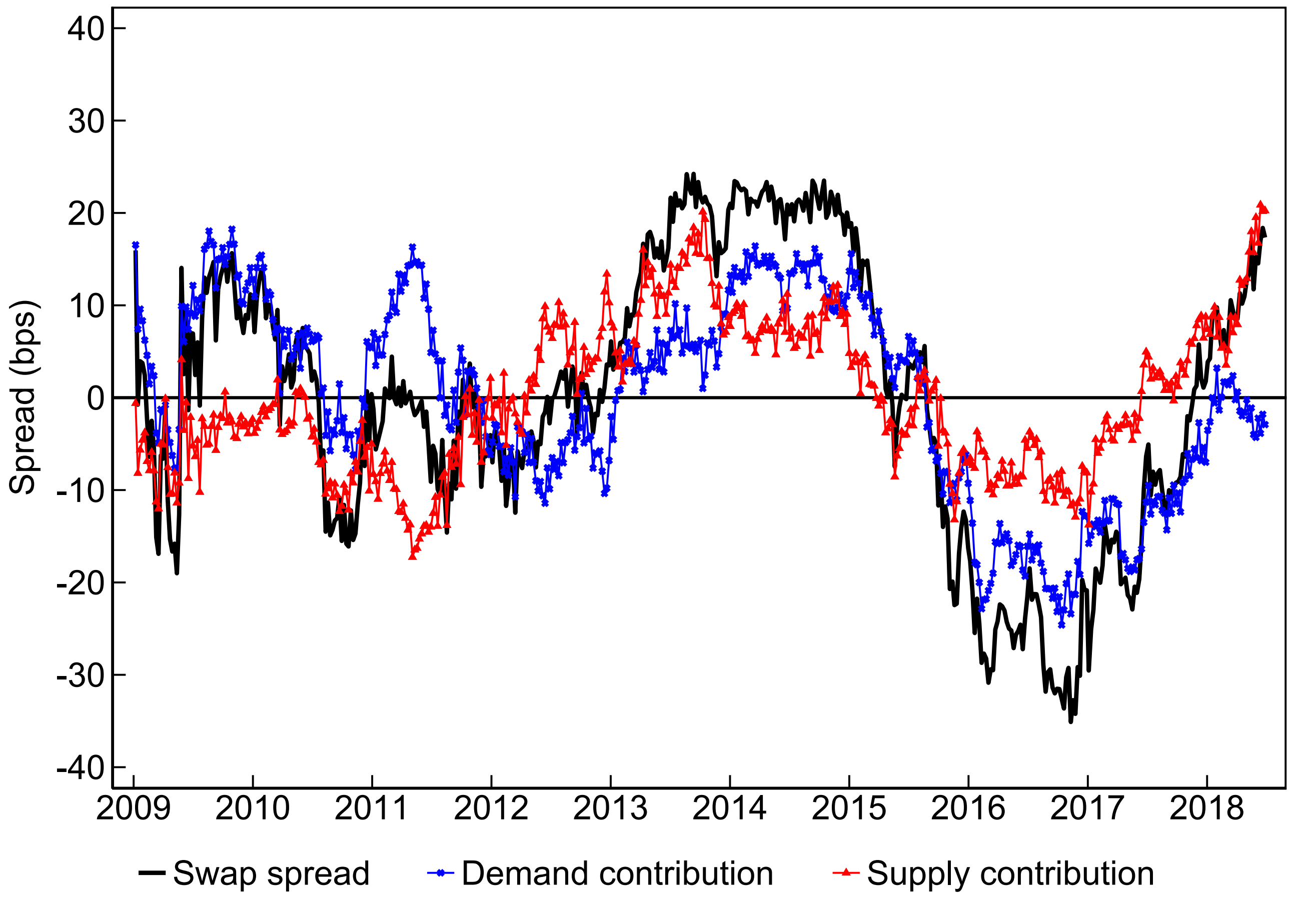
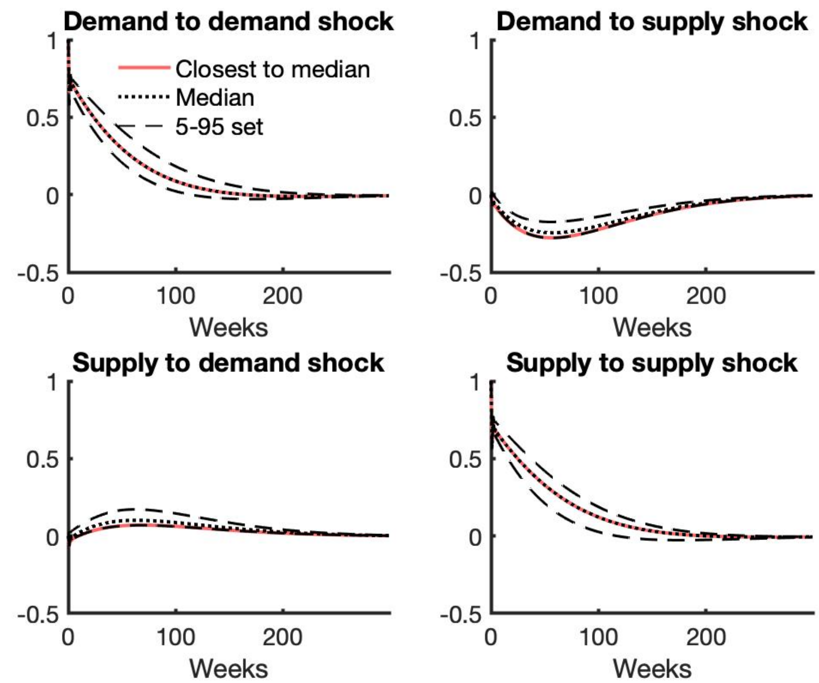


Figure 3: **Swap spread historical decomposition:** This figure shows the 30-year swap spread minus its sample average (Swap spread), and the respective contributions of the demand factor (Demand contribution) and the supply factor (Supply contribution) in $s_t^{(30)} - \mathrm{E}[s_t^{(30)}] = \hat{A}_d \hat{z}_t^d + \hat{A}_w \hat{z}_t^w$. The underlying data are weekly and run from January 2009 to June 2018.

Panel A: Impulse-responses for latent demand and supply factors



Panel B: Impulse-responses for observables

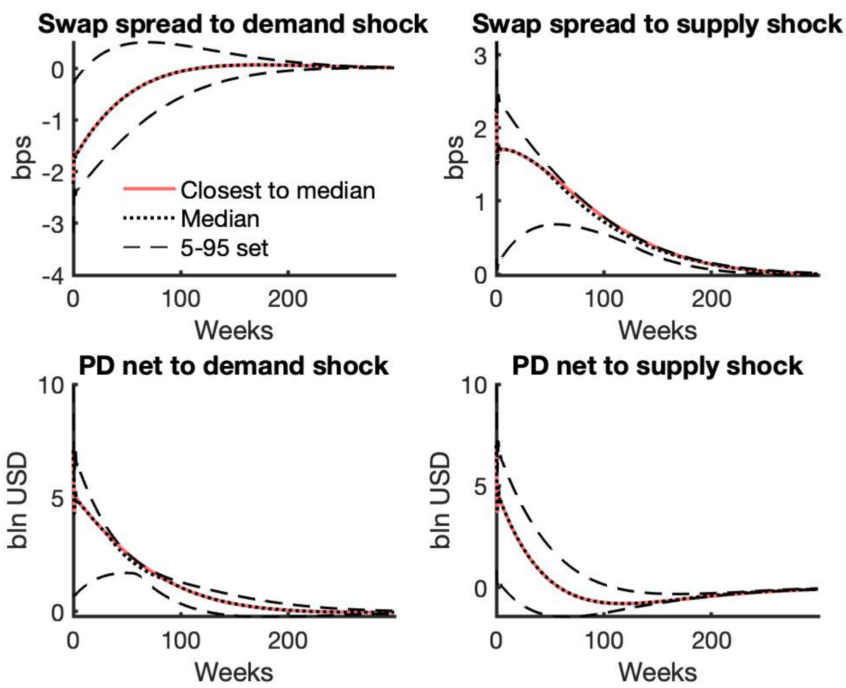


Figure 4: **Impulse response functions for latent demand and supply factors and for observables**: These figures show the impulse response functions from our structural VAR. Median, "closest to median," and 5$^{\text{th}}$to 95$^{\text{th}}$ percentiles show the set of impulses responses that satisfy the identification restrictions. Data are weekly and run from January 2009 to June 2018. Panel A shows the responses for the latent demand factor (z_t^d) and the latent supply factor (z_t^s). Panel B shows the responses for 30-year swap spreads ($s_t^{(30)}$) and quantities ($\hat{d}_t = \textit{PD-UST-Net}_t$).

A Proofs and derivations for the baseline model

Proof of Theorem 1. We write the conjectured form of swap spreads from (16) as

$$s_t = A_0 + \mathbf{a}_s' \mathbf{z}_t,$$

where $\mathbf{a}_s = [A_m, A_d, A_w]'$. We can also generalize our affine equilibrium to a generic VAR(1) data-generating process $\mathbf{z}_{t+1} = \boldsymbol{\varrho}\mathbf{z}_t + \boldsymbol{\varepsilon}_{t+1}$, where $\boldsymbol{\varrho}$ and $\text{Var}_t[\boldsymbol{\varepsilon}_{t+1}] = \boldsymbol{\Sigma}$ are potentially non-diagonal. Doing so yields the following fixed-point condition for $\mathbf{a}_s$:

$$\mathbf{a}_s = \overbrace{\left[(1-\delta)^{-1}(\mathbf{I}-\delta\boldsymbol{\varrho}') + \alpha\gamma\left(\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 + V(\mathbf{a}_s)\right)\mathbf{I}\right]^{-1}\left[\mathbf{e}_m - \kappa_x \frac{\alpha\sigma_o^2}{\kappa_o^2}(\kappa_x \mathbf{e}_d - \mathbf{e}_w) - \alpha V(\mathbf{a}_s)\mathbf{e}_d\right]}^{\mathbf{F}(\mathbf{a}_s)}, \tag{A-1}$$

where $\mathbf{I}$ denotes the 3×3 identity matrix, $\mathbf{e}_m = (1,0,0)'$, $\mathbf{e}_d = (0,1,0)'$, $\mathbf{e}_w = (0,0,1)'$, and

$$V(\mathbf{a}_s) = \left(\frac{\delta}{1-\delta}\right)^2 \mathbf{a}_s' \boldsymbol{\Sigma} \mathbf{a}_s \tag{A-2}$$

is the conditional variance of s_t. Letting $\mathbf{a}_s^* = \mathbf{F}(\mathbf{a}_s^*)$ denote a solution to this fixed-point problem, we then have

$$A_0^* = \frac{\overline{m} - \kappa_x \frac{\alpha\sigma_o^2}{\kappa_o^2}\left(\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x \overline{d} - \overline{w}\right) - \alpha V(\mathbf{a}_s^*)\,\overline{d}}{1 + \alpha\gamma\left[\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 + V(\mathbf{a}_s^*)\right]}. \tag{A-3}$$

In the special case where $\boldsymbol{\varrho}$ and $\boldsymbol{\Sigma}$ are diagonal, equations (A-1) and (A-3) simplify to the expressions given in (18b)-(19).

To understand the relevant cases for equilibrium existence and multiplicity, we rewrite (19) to emphasize the dependence of the A_i coefficients given in (18b) on (18d) on V:

$$V = F(V) \equiv \left(\frac{\delta}{1-\delta}\right)^2 \left[\sigma_d^2 (A_d[V])^2 + \sigma_m^2 (A_m[V])^2 + \sigma_w^2 (A_w[V])^2\right]. \tag{A-4}$$

Thus, the system of equations $\mathbf{a}_s^* = \mathbf{F}(\mathbf{a}_s^*)$ collapses to a scalar fixed-point problem in $V \geq 0$. An equilibrium is stable if it is robust to small perturbations in intermediaries' beliefs about volatility. Formally, an equilibrium with volatility $V^* = F(V^*)$ is stable if and only if $F'(V^*) < 1$—i.e., if the function $F(V)$ crosses the 45-degree line from above at $V = V^*$.

We first consider the case where $\gamma = 0$. When $\gamma = 0$, $A_m[V]$ and $A_w[V]$ are positive constants. If, in addition, $\alpha\sigma_d^2 = 0$, then $\sigma_d^2 (A_d[V])^2 = 0$, so $F(V)$ is constant, and there exists exactly one valid solution to (A-4) and it is stable. However, when $\alpha\sigma_d^2 > 0$, $|A_d[V]|$ is linearly increasing in V, with $\lim_{V\to 0}|A_d[V]| > 0$ and $\lim_{V\to\infty}|A_d[V]| = \infty$, so $F(V)$ is a quadratic function of V. In this case, it is easy to show that (i) there are either zero or two real roots both of which are non-negative, (ii) the two real roots exist if and only if intermediary risk aversion α is below a threshold $\alpha^* > 0$, and (iii) the solution with lower volatility is always stable (i.e., $F'(V^*) < 1$ at this solution) whereas the solution with higher volatility equilibrium is unstable (i.e., $F'(V^*) > 1$ at this solution).

Staying with the case where $\gamma = 0$ and $\alpha\sigma_d^2 > 0$, we next discuss the equilibrium $\mathbf{a}_s^*$ coefficients in a few special cases to better understand the intuition. In particular:

1. If $\kappa_x = 0$ and $m_t = \sigma_m = 0$, we have $A_m^* = A_d^* = A_w^* = 0$ in the unique stable equilibrium. However, there is an unstable equilibrium with $A_d^* < 0$. This is analogous to the unstable equilibrium in De Long et al. (1990) where LoOP fails because all arbitrageurs think it will fail in the future. However, this equilibrium is not robust to small perturbations in intermediary beliefs.

2. If $\kappa_x = 0$ and $\sigma_m > 0$, we have $A_m^* > 0$, $A_d^* < 0$, and $A_w^* = 0$ in the unique stable equilibrium. This is just a standard Vayanos and Vila (2021)-style equilibrium with random net supply and a risky fundamental (i.e., the short-rate differential).

3. If $\kappa_x > 0$, $\sigma_w > 0$, and $m_t = \sigma_m = 0$, we have $A_m^* = 0$, $A_d^* < 0$, and $A_w^* > 0$ in the unique stable equilibrium. This is just a standard Vayanos and Vila (2021)-style equilibrium with random supply and a stochastic shadow value of intermediary capital which functions like a risky fundamental.

4. Finally, if $\kappa_x > 0$, $\sigma_w > 0$, and $\sigma_m > 0$, we have $A_m^* > 0$, $A_d^* < 0$, and $A_w^* > 0$ in the unique stable equilibrium.

We next consider the case where $\gamma > 0$. When $\gamma > 0$, $A_m[V]$ and $A_w[V]$ are decreasing and have finite limits when $V \to 0$ and $V \to \infty$. In particular, $\lim_{V\to 0} A_m[V]$, $\lim_{V\to 0} A_w[V] > 0$, and $\lim_{V\to\infty} A_m[V] = \lim_{V\to\infty} A_w[V] = 0$. On the other hand, $|A_d[V]|$ is increasing in V and has limits $\lim_{V\to 0}|A_d[V]| > 0$ and $\lim_{V\to\infty}|A_d[V]| = 1/\gamma < \infty$. Thus, the RHS of (A-4) must cross the 45-degree line at least once from above as we increase V from zero to ∞. Therefore, when $\gamma > 0$ a stable equilibrium in which $V \geq 0$ always exists.

If $\gamma > 0$ and $\alpha\sigma_d^2 = 0$, this solution is unique. When $\gamma > 0$ and $\alpha\sigma_d^2 > 0$, however, we can have multiple equilibria as above. In fact, when we also have $\kappa_x > 0$, $\sigma_w^2 > 0$, and $\sigma_m^2 > 0$, the fixed-point problem (A-4) is equivalent to finding the roots of a 7th-order polynomial in V. (For SOFR swaps, where $\sigma_m^2 = 0$, this reduces to a 5th-order polynomial). We have shown above that this polynomial always has at least one positive real root that is stable. And, while there can be multiple solutions that are positive and stable in this case, all solutions but the smallest solution diverge in the limit where $\sigma_d^2 \to 0$—i.e., we have $V^* \to \infty$ as $\sigma_d^2 \to 0$ for these solutions. By contrast, the smallest solution remains finite in the limit where $\sigma_d^2 \to 0$. ■

Proof of Proposition 1. Let $\overline{m}^* = \kappa_x\left(\left(\alpha\sigma_o^2\right)/\kappa_o^2\right)\left[\kappa_o\left(\overline{r}_o/\left(\alpha\sigma_o^2\right)\right) + \kappa_x\overline{d} - \overline{w}\right] + \alpha V\overline{d}$. From (18a), it is immediate that $A_0 < 0$ as long as $\overline{m} < \overline{m}^*$. ■

Proof of Proposition 2. From (10), market clearing, and assumption A1, we have $\mathrm{E}_t[r_{t+1}^{Spread}] = -\kappa_x \cdot \psi_t - \alpha V \cdot d_t$. The expression for ψ_t follows from (13) and assumption A2. ■

Proof of Proposition 3. The results on the B_i, $i = \{m, d, w\}$, coefficients follow directly from the the definition of swap returns and the equilibrium coefficients (18b)-(18a). ■

Proof of Proposition 4. Equation (21) follows from the definition of m_t, (7), and (16). Equation (22) follows from combining (21) with (8). ■

Internet Appendix for "Demand-and-Supply Imbalance Risk and Long-Term Swap Spreads"

Not for publication

B Additional theoretical results

B.1 Swap spreads and the outside investment opportunity

Swap intermediaries allocate their scarce capital between the swap spread trade and a risky outside investment opportunity—e.g., another relative-value fixed-income trade. As a result, the equilibrium shadow value of intermediary capital and, hence, swap spreads depend on the expected return and risk of this outside investment opportunity. In particular, we have:

Proposition 7 *Equilibrium swap spreads depend on the expected return ($\bar{r}_o$) of intermediaries' outside investment opportunity. In particular, when $d_t > 0$ we have $\partial E[s_t]/\partial \bar{r}_o = -\kappa_x \cdot \partial E[\psi_t]/\partial \bar{r}_o < 0$, implying that swap spreads s_t are decreasing in $\bar{r}_o$. When $\gamma > 0$, the scale of intermediaries' positions $|x_t| = |d_t|$ is also decreasing in $\bar{r}_o$.*

Proof. From (18a), $\partial \text{E}[s_t]/\partial \bar{r}_o = -\kappa_x \cdot \partial \text{E}[\psi_t]/\partial \bar{r}_o = \partial A_0/\partial \bar{r}_o < 0$ since an increase in $\bar{r}_o$ lowers the numerator of A_0 but does not affect the denominator. When $\gamma > 0$, we have $|\text{E}[x_t]| = |\text{E}[d_t]| = \bar{d} + \gamma A_0$, which is also decreasing in $\bar{r}_o$. ■

Intuitively, the average shadow value of intermediary capital is increasing in $\bar{r}_o$—that is, $\partial \text{E}[\psi_t]/\partial \bar{r}_o > 0$. Thus, when intermediaries are short swap spreads in equilibrium, a rise in $\bar{r}_o$ must be associated with a decline in $\text{E}[s_t] = A_0$. This means that fluctuations in the expected returns on the outside investment opportunity will induce supply shocks similar to those due to fluctuations in the amount of intermediary capital w_t.[29] This link between swap spreads and $\bar{r}_o$ does not rely on the idea that swap intermediaries—the marginal investors in swap market—are also marginal in the outside investment opportunity. In this way, our model has a segmented-markets flavor and contrasts with more integrated-market models where a single set of intermediaries is marginal in a large number of asset classes (He and Krishnamurthy, 2013; Adrian et al., 2014; He et al., 2017).

B.2 Alternate specification for the outside investment opportunity

Assume that instead of the risky outside investment opportunity with expected return $\bar{r}_o$ and variance σ_o^2, intermediaries have access to a riskless outside arbitrage opportunity with a certain 1-period return. However, we assume that this riskless arbitrage return is decreasing in the amount of swap intermediary capital committed to this outside opportunity. As before we assume that swap intermediaries are subject to a leverage constraint: the more intermediaries invest in the swap spread trade the less they can invest in the riskless outside arbitrage opportunity.

Intermediaries' optimization problem can be written as

$$\max_{x_t,\, a_t} \left\{ \text{E}_t\left[w_{t,\,t+1}\right] - \frac{\alpha}{2}\text{Var}_t\left[w_{t,\,t+1}\right] \right\}, \tag{A-5}$$

subject to the budget constraint

$$w_{t,\,t+1} = w_t + x_t r^s_{t+1} + o_t r^o_{t+1} \tag{A-6}$$

and the leverage constraint

$$\kappa_x\,|x_t| + \kappa_o\,|o_t| \leq w_t, \tag{A-7}$$

where r^s_{t+1} is given by (3) while

$$r^o_{t+1} \equiv \bar{r}_o - \lambda_o o_t, \tag{A-8}$$

with $\bar{r}_o, \lambda_o > 0$. This latter assumption can be microfounded by assuming that intermediaries exploit an arbitrage opportunity across two segmented markets and that their position in the long-short arbitrage affects the return on the arbitrage strategy in the spirit of Gromb and Vayanos (2002) and Kondor (2009).

[29] The model can be easily extended so that fluctuations in ψ_t stem from exogenous movements in $\bar{r}_{o,t} = \text{E}_t[r^o_{t+1}]$.

As before, letting $\psi_t \geq 0$ denote the Lagrange multiplier associated with the leverage constraint (A-7), intermediaries' first-order condition for x_t is

$$\mathrm{E}_t[r^s_{t+1}] = \kappa_x sgn\left(x_t\right) \cdot \psi_t + \alpha V_t \cdot x_t, \tag{A-9}$$

where $V_t \equiv \mathrm{Var}_t[r^s_{t+1}] = \left(\frac{\delta}{1-\delta}\right)^2 \mathrm{Var}_t[s_{t+1}]$ is the conditional variance of r^s_{t+1}. Similarly, the first order condition for o_t is

$$r^o_{t+1} = \bar{r}_o - \lambda_o o_t = \kappa_o sgn\left(o_t\right) \cdot \psi_t. \tag{A-10}$$

Since $\bar{r}_o > 0$ and $\psi_t \geq 0$, we must have $o_t \geq 0$, implying that the Lagrange multiplier is

$$0 \leq \psi_t = \frac{\lambda_o}{\kappa_o}\left(\frac{\bar{r}_o}{\lambda_o} - o_t\right) = \max\left\{0, \frac{\lambda_o}{\kappa_o^2}\left(\kappa_o \frac{\bar{r}_o}{\lambda_o} + \kappa_x \left|x_t\right| - w_t\right)\right\}. \tag{A-11}$$

This is isomorphic to the expression in our baseline model setting $\alpha\sigma_o^2 = \lambda_o$. In sum, there are alternative microfoundations that yield similar specifications for intermediaries' supply of swaps.

B.3 General (potentially) nonlinear model

B.3.1 Summary

In this Appendix, we refine our general (potentially) nonlinear characterization of swap spreads. For simplicity, we focus on the case where swaps have zero fundamental value ($m_t \equiv 0$) and where end-user demand for swaps is inelastic ($\gamma = 0$), so there are just two exogenous state variables: end-user demand (d_t) to receive the fixed swap rate and intermediary wealth (w_t). We let $s_t = s\left(d_t, w_t\right)$ denote the equilibrium level of swap spreads,

$$\mathrm{E}_t[r^s_{t+1}] = s_t - \frac{\delta}{1-\delta}\left(\mathrm{E}_t\left[s_{t+1}\right] - s_t\right)$$

the expected return on the receive-fixed swap spread trade, and

$$V(d_t, w_t) \equiv \mathrm{Var}_t[r^s_{t+1}] = \left(\frac{\delta}{1-\delta}\right)^2 \mathrm{Var}_t[s_{t+1}] \tag{A-12}$$

the conditional variance of the returns on this trade.[30]

Two features of our general model can give rise to nonlinearity. First, intermediaries' balance sheet constraint is itself nonlinear, depending on $|d_t|$ instead of d_t: this is relevant if is possibilite that end-user demand—and hence the balance sheet cost of intermediating swaps—may switch signs. Second, intermediaries' balance sheet constraint may not always bind. This means that the *risk* of the swap spread trade is time-varying. The affine equilibrium that we emphasize in the main text arises when we make simplifying assumptions that switch off both of these features, namely, that (A1) end-user demand is always positive and (A2) intermediaries' balance-sheet constraint is always binding. Formally, it turns out that much of the complexity of the nonlinear model that results from relaxing these assumptions stems from the nonlinear nature of the constraint—i.e., it is quite straightforward to relax A2, but relaxing A1 is somewhat more involved.

In this Appendix, we characterize a number of important special cases of this (potentially) nonlinear model:

- *Case 1*: When both A1 and A2 hold, we obtain 1 above. Specifically, s_t and $\mathrm{E}_t[r^s_{t+1}]$ are affine functions of d_t and w_t and $V_t = V$ is constant. Furthermore, we have $\partial s_t/\partial d_t = A_d < 0$ and $\partial \mathrm{E}_t[r^s_{t+1}]/\partial d_t \propto A_d < 0$, and $\partial s_t/\partial w_t = A_w > 0$ and $\partial \mathrm{E}_t[r^s_{t+1}]/\partial w_t \propto A_w > 0$.
- *Case 2*: When we relax A2 but maintain A1—i.e., if intermediaries' balance-sheet constraint can be slack, but end-user demand is always positive—then s_t, $\mathrm{E}_t\left[r^s_{t+1}\right]$, and V_t are non-linear functions that satisfy $\partial s_t/\partial d_t < 0$ and $\partial s_t/\partial w_t > 0$; $\partial \mathrm{E}_t\left[r^s_{t+1}\right]/\partial d_t < 0$ and $\partial \mathrm{E}\left[r^s_{t+1}\right]/\partial w_t > 0$; and, finally, $\partial V_t/\partial d_t > 0$ and $\partial V_t/\partial w_t < 0$.

  Intuitively, even if leverage constraints are not binding at time t, the mere potential for them to bind in the future makes swap spread trades risky for intermediaries and they will only accommodate end-user demand for long-term swaps if they are compensated for this risk (i.e., we have $V_t > 0$ even when $\psi_t = 0$). Furthermore, even if the constraint is currently slack, fluctuations in end-user demand and intermediary capital will shape the likelihood that the constraint will bind in the future and hence the current level of spreads.

  Thus, the only qualitative difference between *Case 2* and the affine model of *Case 1* is that the risk of the swap spread trade now fluctuates endogenously as a function of d_t and w_t. Specifically, swap spreads become more

[30]We sometimes use the shorthand notation $\mathrm{E}_t\left[\cdot\right] \equiv \mathrm{E}[\cdot|d_t, w_t]$ and $\mathrm{Var}_t\left[\cdot\right] \equiv \mathrm{Var}[\cdot|d_t, w_t]$.

volatile when it is more likely that intermediary constraints will bind in the future, explaining why the variance is increasing in end-user demand (d_t) and decreasing in intermediary wealth (w_t). This endogenous-risk effect adds a further channel through which increases in current end-user demand (or reductions in intermediary wealth) raise the expected returns on the swap spread trade and, hence, the current level of swap spreads.

- *Case 3*: The solution can also be characterized rigorously if we relax both A1 and A2, but make the model symmetric by assuming that the steady-state level of end-user demand is zero ($\overline{d} = 0$). When $\overline{d} = 0$, the sign of demand may switch over time, but the expected sign of future demand is always the same as the sign of current demand, making the model quite tractable. In this case, s_t, $\mathrm{E}_t\left[r^s_{t+1}\right]$, and V_t are non-linear functions that satisfy $\partial s_t/\partial d_t < 0$ and $sgn\left(\partial s_t/\partial w_t\right) = sgn\left(d_t\right)$; $\partial \mathrm{E}_t\left[r^s_{t+1}\right]/\partial d_t < 0$ and $sgn\left(\partial \mathrm{E}\left[r^s_{t+1}\right]/\partial w_t\right) = sgn\left(d_t\right)$; and, finally, $sgn\left(\partial V_t/\partial d_t\right) = sgn\left(d_t\right)$ and $\partial V_t/\partial w_t < 0$.

  Relative to *Case 2*, there is now an additional source of non-linearity due to the fact that end-user demand can switch sign over time. This means that $sgn\left(s_t\right) = sgn\left(\mathrm{E}_t\left[r^s_{t+1}\right]\right) = -sgn\left(d_t\right)$ and explains why the impact of intermediary wealth on expected returns and spreads depends on the sign of end-user demand. Furthermore, as in *Case 2* above, swap spreads become more volatile when it is more likely that constraints will bind in the future. Since the constraint depends on $|d_t|$ and we assume $\overline{d} = 0$, this explains why variance is decreasing in end-user demand when $d_t < 0$ and increasing in end-user demand when $d_t > 0$.

- *Case 4*: We can also rigorously characterize the model in the limit where $\alpha = 0$ for an arbitrary value of $\overline{d} \neq 0$—i.e., for risk-neutral intermediaries who are subject to balance-sheet constraints.[31] This means are relaxing both A1 and A2 as well as introducing potential asymmetries between positive and negative demand. However, there are no risk premia, so swap spreads are just the average expected balance sheet costs over the life of the swap. In this case, we find that s_t, $\mathrm{E}_t\left[r^s_{t+1}\right]$, and V_t are non-linear functions that satisfy $\partial s_t/\partial d_t < 0$, $\partial \mathrm{E}_t\left[r^s_{t+1}\right]/\partial d_t < 0$, $sgn\left(\partial \mathrm{E}_t[r^s_{t+1}]/\partial w_t\right) = sgn\left(d_t\right)$, and $\partial V_t/\partial w_t < 0$. Furthermore, there will exist some $d^*\left(w_t\right)$, whose sign is opposite that of $\overline{d}$, such that $\partial s_t/\partial w_t > 0$ for all $d_t > d^*\left(w_t\right)$ and $\partial s_t/\partial w_t < 0$ for all $d_t < d^*\left(w_t\right)$. Similarly, V_t will be a U-shaped function of d_t as in *Case 3*, although it will now achieve its mininum at some value $d^{**}\left(w_t\right)$ whose sign is opposite that of $\overline{d}$. However, these cutoff thresholds may lie outside of the state space—e.g., when demand is almost always positive, implying that the comparative statics are monotone over the state space.

- Finally, since the model's stable equilibrium is continuous in the underlying model parameters, we can use these special cases to say more about how the model behaves for other parameter values. For instance, if the probability that end-user demand turns negative is sufficiently small, then by continuity the solution must have the same properties as those in *Case 2*. Similarly, if either (i) α is sufficiently small for any value of $\overline{d} \neq 0$ or (ii) if $\left|\overline{d}\right|$ is sufficiently small for any value of $\alpha > 0$, then by continuity the solution must have the same properties as in *Case 4*. And, while our analysis assumes that $\gamma = 0$, this characterization remains valid so long as the demand elasticity γ is not too large.

Overall, the key qualitative implications of our affine model generalize nicely once we allow for nonlinearities. Furthermore, it is clear that the global behavior of $s\left(d_t, w_t\right)$, $\mathrm{E}_t\left[r^s_{t+1}\right]$, and $V(d_t, w_t)$ that we see in these special cases must obtain for any set of model parameters. Indeed, the only thing that we have been unable to rule out—either analytically or numerically—is that the local behavior of $s\left(d_t, w_t\right)$, $\mathrm{E}_t\left[r^s_{t+1}\right]$, and $V(d_t, w_t)$ may conflict with what we see in the special cases. For instance, we have not been able to rule out the possibility that $s\left(d_t, w_t\right)$ could be locally increasing in d_t in some case where α and $\overline{d}$ are large (but not too large since then demand will rarely be negative and the solution will be like that in *Case 2*). However, even if our theory predicted small local monotonicities of this sort for certain parameterizations, we would be inclined to view this as more of a theoretical curiousity than as an empirical prediction with important practical relevance.[32]

B.3.2 Setting

For simplicity, we focus on the case where swaps have zero fundamental value ($m_t \equiv 0$) and where end-user demand for swaps is inelastic ($\gamma = 0$). Thus, there are just two exogenous state variables: intermediary wealth (w_t) and end-user demand (d_t) to receive the fixed swap rate. We assume

$$w_t = \overline{w} + z^w_t, \tag{A-13}$$

where $\overline{w} > 0$ and

$$d_t = \overline{d} + z^d_t. \tag{A-14}$$

[31] To study this limit, we need to modify our micro-foundation for intermediary balance sheet costs to ensure they still vary as a function of d_t and w_t when $\alpha = 0$.

[32] Specifically, any such local non-monotonicities would rely on almost heroic levels of investor foresight and investor knowledge about the underlying data generating processes. And, we believe that treating these as empirical predictions that we should take to the data would be taking our model far too literally.

The sign of $\overline{d}$ can be either positive, negative, or zero.

We assume the vector of mean-zero state variables $\mathbf{z}_t = [z_t^d, z_t^w]'$ follows an AR(1) process

$$\mathbf{z}_{t+1} = \boldsymbol{\varrho}\mathbf{z}_t + \boldsymbol{\varepsilon}_{t+1}, \tag{A-15}$$

where $\boldsymbol{\varrho} = diag(\rho_d, \rho_w)$ is a diagonal matrix of AR(1) coefficients $\rho_d, \rho_w \in [0, 1)$ and $\boldsymbol{\varepsilon}_{t+1} = [\varepsilon_{t+1}^d, \varepsilon_{t+1}^w]'$ is the vector of mean-zero ($\text{E}_t[\boldsymbol{\varepsilon}_{t+1}] = \mathbf{0}$) structural shocks. For simplicity, we assume the two structural shocks are orthogonal to each other and are iid over time: $\text{Var}_t[\boldsymbol{\varepsilon}_{t+1}] = diag\left(\sigma_d^2, \sigma_w^2\right)$, where $\sigma_i^2 \equiv \text{Var}\left[\varepsilon_{t+1}^i\right]$ for $i = w, d$.

In addition, we assume the structural shocks have a bounded support. Formally, for $i = d, w$, we assume $\varepsilon_{t+1}^i \in [-\overline{\varepsilon}^i, \overline{\varepsilon}^i]$ where $\overline{\varepsilon}^i > 0$. Since the shocks have bounded support, the assumed AR(1) dynamics imply that the state variables also have a bounded support. Specifically, for $i = d, w$, we have $z_t^i \in [-\overline{z}^i, \overline{z}^i]$ where $\overline{z}^i \equiv \overline{\varepsilon}^i / (1 - \rho_i)$. This, in turn, implies

$$w_t = \overline{w} + z_t^w \in [\overline{w} - \overline{z}^w, \overline{w} + \overline{z}^w] \tag{A-16}$$

and

$$d_t = \overline{d} + z_t^d \in \left[\overline{d} - \overline{z}^d, \overline{d} + \overline{z}^d\right]. \tag{A-17}$$

Working with a bounded state space allows us to provide sufficient conditions which guarantee that any required parametric restrictions always hold throughout the state space.

B.3.3 General characterization from the main text

In this setting, the general characterization of swap spreads presented in the main text is valid so long as intermediaries always have sufficient capital to accommodate end-user demand (A3). Formally, we only need to assume

$$\kappa_x |d_t| < w_t \tag{A-18}$$

at all points in the state space, and a sufficient condition for this to hold is

$$v \cdot \max\left(\overline{d} + \overline{z}^d, \overline{z}^d - \overline{d}\right) < \overline{w} - \overline{z}^w. \tag{A-19}$$

However, we can allow $sgn(d_t)$ to vary over time and we can allow the balance sheet constraint $\psi_t = \psi(|d_t|, w_t)$ to be slack at some points in the state space. Under this restriction, the expected return on the receive-fixed swap spread trade is

$$\overbrace{s_t - \frac{\delta}{1-\delta}(\text{E}_t[s_{t+1}] - s_t)}^{\text{E}_t[r_{t+1}^s]} = \overbrace{(-\kappa_x) sgn(d_t) \cdot \psi(|d_t|, w_t)}^{\substack{\text{Compensation for using} \\ \text{scarce capital}}} + \overbrace{(-\alpha) V(d_t, w_t) \cdot d_t}^{\substack{\text{Compensation} \\ \text{for risk}}}, \tag{A-20}$$

where

$$\psi(|d_t|, w_t) = \max\left\{0, \frac{\alpha\sigma_o^2}{\kappa_o^2}\left(\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x |d_t| - w_t\right)\right\} \geq 0 \tag{A-21}$$

is the Lagrange multiplier associated with intermediaries' leverage constraint. The level of swap spreads is

$$s_t = (1-\delta)\sum\nolimits_{k=0}^{\infty} \delta^k \text{E}_t\left[(-\kappa_x)\, sgn(d_{t+k})\, \psi(|d_{t+k}|, w_{t+k}) + (-\alpha)\, V(d_{t+k}, w_{t+k}) d_{t+k}\right]. \tag{A-22}$$

In our general model, $V(d_t, w_t)$, $\text{E}_t[r_{t+1}^s]$, and s_t may vary nonlinearly over the state space. This contrasts with our affine equilibrium where $V(d_t, w_t) = V$ is constant and $\text{E}_t[r_{t+1}^s]$ and s_t both vary linearly over the state space.

Formally, the equilibrium level of swap spreads, $s_t = s(d_t, w_t)$, in our non-linear model must solve the following functional fixed-point equation over the state space:

$$s(d_t, w_t) = (1-\delta)\left[\begin{array}{c} (-\kappa_x) sgn(d_t)\psi(|d_t|, w_t) \\ +(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2 \text{Var}\left[s(d_{t+1}, w_{t+1}) | d_t, w_t\right] \cdot d_t \end{array}\right] + \delta \text{E}_t\left[s(d_{t+1}, w_{t+1})\right]. \tag{A-23}$$

Equivalently, the equilibrium variance of the swap spread trade, $V(d_t, w_t)$, defined in (A-12), must solve the following functional fixed-point equation:

$$V(d_t, w_t) = \delta^2 \text{E}_t\left[\left(\sum\nolimits_{k=0}^{\infty} \delta^k (\text{E}_{t+1} - \text{E}_t)\left[\begin{array}{c} \kappa_x sgn(d_{t+1+k})\psi(|d_{t+1+k}|, w_{t+1+k}) \\ +\alpha V(d_{t+1+k}, w_{t+1+k}) d_{t+1+k} \end{array}\right]\right)^2\right]. \tag{A-24}$$

B.3.4 The affine equilibrium

Naturally, the affine model we emphasize in the main text is just a special case of this general model where demand is always positive (A1) and the balance sheet constraint is always binding (A2). Specifically, our affine model obtains if we always have:

$$0 < d_t,\, \kappa_x d_t < w_t, \text{ and } w_t < \kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x d_t. \tag{A-25}$$

The first inequality ($0 < d_t$) says that end-user net demand to receive the fixed swap rate is always positive. The second inequality ($\kappa_x d_t < w_t$) says that intermediaries always have sufficient capital to accommodate end-user demand in the swap market. The third inequality ($w_t < \kappa_o \overline{r}_o / \left(\alpha\sigma_o^2\right) + \kappa_x d_t$) ensures that intermediaries' leverage constraint always binds—i.e., we always have $\psi_t > 0$.

A sufficient condition for equation (A-25) to hold everywhere in the state space is that $\overline{d}$, $\overline{w}$, $\overline{z}^d$, and $\overline{z}^w > 0$ satisfiy:

$$0 < \overline{d} - \overline{z}^d,\, \kappa_x \left(\overline{d} + \overline{z}^d\right) < \left(\overline{w} - \overline{z}^w\right), \text{ and } \left(\overline{w} + \overline{z}^w\right) < \kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x \left(\overline{d} - \overline{z}^d\right). \tag{A-26}$$

So long as $\kappa_o \overline{r}_o / \left(\alpha\sigma_o^2\right) > 0$, we can always choose values of $\overline{d}$, $\overline{w}$, $\overline{z}^d$, and $\overline{z}^w > 0$ so that these three inequalities hold. In this case, the general model reduces to the affine model presented in the main text, and we obtain the following results:[33]

Proposition 8 ***The affine special case.*** *Assume (i) $0 < \overline{d} - \overline{z}^d$ so end-user demand to receive fixed is always positive, (ii) $\kappa_x \left(\overline{d} + \overline{z}^d\right) < \left(\overline{w} - \overline{z}^w\right)$ so intermediaries always have enough capital to accommodate end-user demand, and (iii) $\left(\overline{w} + \overline{z}^w\right) < \kappa_o \overline{r}_o / \left(\alpha\sigma_o^2\right) + \kappa_x \left(\overline{d} - \overline{z}^d\right)$ so intermediaries' constraint is always binding. Then our general model reduces to the affine model in the main text. Specifically, swap spreads are an affine function of the state variables:*

$$s_t = A_0 + A_d z_t^d + A_w z_t^w,$$

where

$$\begin{aligned}
A_0 &= -\kappa_x \frac{\alpha\sigma_o^2}{\kappa_o^2}\left[\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x \overline{d} - \overline{w}\right] - \alpha V \overline{d} < 0, \\
A_d &= -\frac{1-\delta}{1-\rho_d \delta} \cdot \alpha \left[\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 + V\right] < 0, \\
A_w &= \frac{1-\delta}{1-\rho_w \delta} \cdot \left[\frac{1}{\kappa_x} \alpha \left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2\right] > 0,
\end{aligned}$$

and $V = \left(\delta / (1-\delta)\right)^2 \left(A_d^2 \sigma_d^2 + A_w^2 \sigma_w^2\right) > 0$ is the constant equilibrium variance of the swap spread trade. (This is Theorem 1 in the main text.)

The expected return on the received-fixed swap spread trade is:

$$E_t[r_{t+1}^s] = -\kappa_x \cdot \overbrace{\left[\frac{\alpha\sigma_o^2}{\kappa_o^2}\left(\frac{\kappa_o \overline{r}_o}{\alpha\sigma_o^2} + \kappa_x d_t - w_t\right)\right]}^{\psi_t} - \alpha V \cdot d_t$$

which is affine in d_t and w_t, decreasing in d_t, and increasing in w_t (This is Proposition 3.) Furthermore, $E_t[r_{t+1}^s]$ is decreasing in end-user demand d_t, even when controlling for the shadow value of capital ψ_t. (This is Proposition 2.) Finally, both s_t and $E_t[r_{t+1}^s]$ are decreasing in $\overline{r}_o$ (This is Proposition 4.)

B.3.5 Enriching the general characterization

In this section, we enrich our characterization of our general non-linear model. We assume throughout that $\overline{d}$, $\overline{w}$, $\overline{z}^d$, and $\overline{z}^w > 0$ satisfy $k_x \cdot \max\left(\overline{d} + \overline{z}^d, \overline{z}^d - \overline{d}\right) < \overline{w} - \overline{z}^w$, so intermediaries always have sufficient capital to accommodate the exogenous level of end-user demand. However, we assume it is either the case that (i) the sign of end-user demand can fluctuate over time (i.e., we can have $\overline{d} - \overline{z}^d < 0 < \overline{d} + \overline{z}^d$), (ii) intermediaries' balance sheet constraint can be slack (i.e., $\overline{w} + \overline{z}^w > \kappa_o \overline{r}_o / \left(\alpha\sigma_o^2\right) + \kappa_x \min_{z^d \in [-\overline{z}^d, \overline{z}^d]} \left|\overline{d} + z^d\right|$), or both (i) and (ii). Under these conditions, our general model will indeed be non-linear.

[33] Of course, we can also construct a symmetric affine equilibrium we have $d_t < 0$, $\kappa_x |d_t| < w_t$, and $w_t < \kappa_o \overline{r}_o / \left(\alpha\sigma_o^2\right) + \kappa_x |d_t|$ everywhere in the state space.

The limiting case where $\alpha = 0$. To build intuition about the potential sources of non-linearity in our model, we study the limiting case where $\alpha = 0$. When $\alpha = 0$, there are no risk premia and swap spreads are simply the discounted expected value of future balance sheet costs:

$$s(d_t, w_t) = (1-\delta)\sum_{k=0}^{\infty} \delta^k \mathrm{E}_t\left[(-\kappa_x)\, sgn(d_{t+k})\, \psi(|d_{t+k}|, w_{t+k})\right]. \tag{A-28}$$

However, to make the analysis interesting we consider the alternate micro-foundational for the outside investment constraint introduced above. This means that the Lagrange multiplier is given by

$$\psi(|d_t|, w_t) = \max\left\{0, \frac{\lambda_o}{\kappa_o^2}\left(\kappa_o \frac{\overline{r}_o}{\lambda_o} + \kappa_x |x_t| - w_t\right)\right\},$$

where $\lambda_o > 0$. So long as $\lambda_o > 0$, this depends on d_t and w_t even when $\alpha = 0$.[34]

We then want to differentiate equation (A-28) with respect to d_t and w_t. To do so, we use the following lemma:

Lemma 1 *Consider $E_t\left[(-\kappa_x)\, sgn(d_{t+k})\, \psi(|d_{t+k}|, w_{t+k})\right]$—i.e., the time-$t$ expectation of the balance sheet compensation at time $t+k$. We have*

$$\begin{aligned}&\frac{\partial E_t\left[(-\kappa_x)\, sgn(d_{t+k})\, \psi(|d_{t+k}|, w_{t+k})\right]}{\partial d_t} \\ &\quad= -\left(\frac{\kappa_x}{\kappa_o}\right)^2 \lambda_o \cdot \rho_d^k \cdot Pr\left[\psi_{t+k} > 0 | d_t, w_t\right] - 2\kappa_x \rho_d^k \cdot E_t\left[\Delta(d_{t+k})\, \psi(|d_{t+k}|, w_{t+k})\right] < 0,\end{aligned}$$

where $\Delta(x) = \lim_{\sigma\to 0} \sigma^{-1}\phi(x/\sigma)$ is Dirac's delta function and $\phi(\cdot)$ is the pdf of the standard normal distribution. Moreover,

$$\frac{\partial E_t\left[(-\kappa_x)\, sgn(d_{t+k})\, \psi(|d_{t+k}|, w_{t+k})\right]}{\partial w_t} = \frac{1}{\kappa_x}\left(\frac{\kappa_x}{\kappa_o}\right)^2 \lambda_o \cdot \rho_w^k \cdot \left(\begin{array}{c} Pr\left[\psi_{t+k} > 0, d_{t+k} > 0 | d_t, w_t\right] \\ -Pr\left[\psi_{t+k} > 0, d_{t+k} < 0 | d_t, w_t\right]\end{array}\right).$$

Proof. Note that

$$\begin{aligned}&\mathrm{E}_t\left[(-\kappa_x)\, sgn(d_{t+k})\, \psi(|d_{t+k}|, w_{t+k})\right] \\ &\quad= \mathrm{E}_t\left[(-\kappa_x)\, sgn\left(\overline{d} + \rho_d^k\left(d_t - \overline{d}\right) + \varepsilon_{t\to t+k}^d\right) \psi\left(\left|\overline{d} + \rho_d^k\left(d_t - \overline{d}\right) + \varepsilon_{t\to t+k}^d\right|, \overline{w} + \rho_w^k(w_t - \overline{w}) + \epsilon_{t\to t+k}^w\right)\right],\end{aligned}$$

where the expectation is taken over the distribution of $\varepsilon_{t\to t+k}^d = \sum_{j=1}^k (\rho_d)^{k-j} \varepsilon_{t+k}^d \in \left[-\overline{\varepsilon}^d, \overline{\varepsilon}^d\right]$ and $\epsilon_{t\to t+k}^w = \sum_{j=1}^k (\rho_w)^{k-j} \varepsilon_{t+k}^w \in \left[-\overline{\varepsilon}^w, \overline{\varepsilon}^w\right]$—i.e., the cumulative innovation to d_{t+k} and w_{t+k} from t to $t+k$. Thus, d_t shows up twice inside the relevant integrand: once inside $\psi(|d_{t+k}|, w_{t+k})$ and once inside $sgn(d_{t+k})$. The latter appearance is only relevant if $sgn(d_{t+k})$ is random.

Recalling that $sgn(d_{t+k}) = \lim_{\sigma\to 0}(2\Phi(d_{t+k}/\sigma) - 1)$, where $\Phi(\cdot)$ is the cdf of the standard normal distribution, we have

$$\begin{aligned}&\frac{\partial}{\partial d_t}\mathrm{E}_t\left[(-\kappa_x)\, sgn(d_{t+k})\, \psi(|d_{t+k}|, w_{t+k})\right] = \lim_{\sigma\to 0} \mathrm{E}_t\left[\frac{\partial}{\partial d_t}(-\kappa_x)\left(2\Phi\left(\frac{d_{t+k}}{\sigma}\right) - 1\right)\psi(|d_{t+k}|, w_{t+k})\right] \\ &\quad= \lim_{\sigma\to 0} \mathrm{E}_t\left[(-\kappa_x)\left(2\Phi\left(\frac{d_{t+k}}{\sigma}\right) - 1\right) \cdot \frac{\partial \psi(|d_{t+k}|, w_{t+k})}{\partial |d_{t+k}|}\frac{\partial |d_{t+k}|}{\partial d_{t+k}}\frac{\partial d_{t+k}}{\partial d_t}\right] \\ &\qquad+ \lim_{\sigma\to 0} \mathrm{E}_t\left[(-\kappa_x)\frac{2}{\sigma}\phi\left(\frac{d_{t+k}}{\sigma}\right)\frac{\partial d_{t+k}}{\partial d_t} \cdot \psi(|d_{t+k}|, w_{t+k})\right] \\ &\quad= \mathrm{E}_t\left[(-\kappa_x)\, sgn(d_{t+k}) \cdot \frac{\partial \psi(|d_{t+k}|, w_{t+k})}{\partial |d_{t+k}|}\rho_d^k sgn(d_{t+k})\right] + \mathrm{E}_t\left[(-2\kappa_x)\,\Delta(d_{t+k})\,\rho_d^k \cdot \psi(|d_{t+k}|, w_{t+k})\right] \\ &\quad= -\left(\frac{\kappa_x}{\kappa_o}\right)^2 \lambda_o \cdot \rho_d^k \cdot \mathrm{E}_t\left[\mathbb{I}\left\{\psi_{t+k} > 0\right\}\right] - 2\kappa_x \cdot \rho_d^k \cdot \mathrm{E}_t\left[\Delta(d_{t+k})\, \psi(|d_{t+k}|, w_{t+k})\right] \\ &\quad= -\left(\frac{\kappa_x}{\kappa_o}\right)^2 \lambda_o \cdot \rho_d^k \cdot \mathrm{Pr}\left[\psi_{t+k} > 0 | d_t, w_t\right] - 2\kappa_x \cdot \rho_d^k \cdot \mathrm{E}_t\left[\Delta(d_{t+k})\, \psi(|d_{t+k}|, w_{t+k})\right] < 0,\end{aligned}$$

[34] By contrast, our baseline formulation imposes $\lambda_o = \alpha\sigma_0^2$, so $\psi(|d_t|, w_t) = \overline{r}_o/\kappa_o$ is constant in the limit where $\alpha = 0$.

where $\mathbb{I}(\cdot)$ is the indicator function.[35]

Assuming that $sgn(d_{t+k})$ is fixed, an increase in the level of end-user demand at time t reduces the expected future compensation for using scarce balance sheet at $t+k$ in proportion to the conditional probability that the balance sheet constraint binds at $t+k$ (i.e., $\Pr[\psi_{t+k} > 0|d_t, w_t]$).[36] The second term involving Dirac's delta function, $\mathrm{E}[\Delta(d_{t+k})\psi(|d_{t+k}|, w_{t+k})|d_t, w_t]$, arises only when $sgn(d_{t+k})$ is stochastic and accounts for the fact that a higher current level of d_t increases expected future balance-sheet costs because it leads $sgn(d_{t+k})$ to switch from -1 to 1 in some future states of the world.[37]

A similar calculation shows that:

$$\frac{\partial \mathrm{E}_t[(-\kappa_x) sgn(d_{t+k})\psi(|d_{t+k}|, w_{t+k})]}{\partial w_t} = \frac{1}{\kappa_x}\cdot\lambda_o\left(\frac{\kappa_x}{\kappa_o}\right)^2\cdot\rho_w^k\cdot\mathrm{E}_t\left[sgn(d_{t+k})\mathbb{I}\left(\psi_{t+k}>0\right)\right]$$
$$= \frac{1}{\kappa_x}\cdot\lambda_o\left(\frac{\kappa_x}{\kappa_o}\right)^2\cdot\rho_w^k\cdot\left(\Pr\left[\psi_{t+k}>0, d_{t+k}>0|d_t, w_t\right] - \Pr\left[\psi_{t+k}>0, d_{t+k}<0|d_t, w_t\right]\right).$$

Thus, an increase in the current level of wealth has a potentially ambiguous impact on the expected future compensation for using scarce balance sheet at $t+k$ ($\mathrm{E}_t[(-\kappa_x) sgn(d_{t+k})\psi(|d_{t+k}|, w_{t+k})]$). This ambiguity arises since the effect of a change in intermediary wealth depends on the likelihood that the constraint will bind in the future and the likely sign of future end-user demand. ■

Using this Lemma we can compute the impact of a change in d_t on swap spreads. We obtain:

$$\frac{\partial s(d_t, w_t)}{\partial d_t} = -\left(\frac{\kappa_x}{\kappa_o}\right)^2\lambda_o\frac{1-\delta}{1-\delta\rho_d}\cdot\overbrace{(1-\delta\rho_d)\sum\nolimits_{k=0}^{\infty}(\delta\rho_d)^k\Pr\left[\psi_{t+k}>0|d_t, w_t\right]}^{\mathcal{P}^d_{\psi>0}(d_t,w_t)\in(0,1)} \tag{A-29}$$
$$-2\kappa_x\frac{1-\delta}{1-\delta\rho_d}\cdot\underbrace{(1-\delta\rho_d)\sum\nolimits_{k=0}^{\infty}(\delta\rho_d)^k\mathrm{E}_t\left[\Delta(d_{t+k})\psi(|d_{t+k}|, w_{t+k})\right]}_{\Delta_{\psi,d=0}(d_t,w_t)>0}<0,$$

where $\mathcal{P}^d_{\psi>0}(d_t, w_t)$ is an averaged probability that the constraint binds over the future life of the swap and $\Delta_{\psi,d=0}(w_t,, d_t)$ captures the effect of higher levels of d_t on the future sign of d_{t+k}.

In our affine model, we always have $\mathcal{P}^d_{\psi>0}(d_t, w_t) = 1$ and $\Delta_{\psi,d=0}(w_t,, d_t) = 0$, so changes in end-user demand have a linear impact on swap spreads. In our general model, $\mathcal{P}^d_{\psi>0}(d_t, w_t) \in (0,1)$ and $\Delta_{\psi,d=0}(w_t,, d_t) > 0$ will vary over the state space, so changes in end-user demand will have a non-linear impact on spreads.

We next differentiate equation (A-28) with respect to w_t. Using our lemma, we obtain

$$\frac{\partial s(d_t, w_t)}{\partial w_t} = \frac{1}{\kappa_x}\left(\frac{\kappa_x}{\kappa_o}\right)^2\lambda_o\frac{1-\delta}{1-\delta\rho_w}\cdot\overbrace{\left\{(1-\delta\rho_w)\sum\nolimits_{k=0}^{\infty}(\delta\rho_w)^k\begin{pmatrix}\Pr\left[\psi_{t+k}>0, d_{t+k}>0|d_t, w_t\right]\\ -\Pr\left[\psi_{t+k}>0, d_{t+k}<0|d_t, w_t\right]\end{pmatrix}\right\}}^{\mathcal{P}^w_{\psi>0,d>0}(d_t,w_t)-\mathcal{P}^w_{\psi>0,d<0}(d_t,w_t)\in(-1,1)}. \tag{A-30}$$

The term in curly braces, $\mathcal{P}^w_{\psi>0,d>0}(d_t, w_t) - \mathcal{P}^w_{\psi>0,d<0}(d_t, w_t) \in (-1,1)$, is an averaged probability that the constraint binds with positive demand over the life of the swap minus the corresponding averaged probability that the constraint binds with negative demand.[38]

---

[35]Note that

$$\mathrm{E}_t[\Delta(d_{t+k})\psi(w_{t+k}, |d_{t+k}|)] = \int\int\Delta(d_{t+k})\psi(w_{t+k}, |d_{t+k}|)f_w(w_{t+k}|w_t)f_d(d_{t+k}|d_t)\,\mathrm{d}w_{t+k}\mathrm{d}d_{t+k}$$
$$= \int\left[\int\psi(w_{t+k}, |d_{t+k}|)f_d(d_{t+k}|d_t)\Delta(d_{t+k})\,\mathrm{d}d_{t+k}\right]f_w(w_{t+k}|w_t)\,\mathrm{d}w_{t+k}$$
$$= \int\left[\psi(w_{t+k}, 0)f_d(0|d_t)\right]f_w(w_{t+k}|w_t)\,\mathrm{d}w_{t+k} = \mathrm{E}\left[\psi(w_{t+k}, 0)f_d(0|d_t)|w_t\right].$$

[36]Technically, the derivatives of $sgn(d_t)\psi(w_t, |d_t|)$ with respect to w_t and d_t do not exist on the boundary where the leverage constraint binds exactly (i.e., where $\kappa_o\overline{r}_o/\left(\alpha\sigma_o^2\right) + \kappa_x|d_t| - w_t = 0$). Since $sgn(d_t)\psi(w_t, |d_t|) = 0$ on this boundary, this does not impact this expectation.

[37]One can see that this term is not ignorable by considering the special case where the constraint always binds but where demand can switch sign.

[38]Since $\mathcal{P}^w_{\psi>0,d>0}(w_t, d_t) + \mathcal{P}^w_{\psi>0,d<0}(w_t, d_t) = \mathcal{P}^w_{\psi>0}(w_t, d_t)$, we can also write $\mathcal{P}^w_{\psi>0,d>0}(w_t, d_t) - \mathcal{P}^w_{\psi>0,d<0}(w_t, d_t) = 2\mathcal{P}^w_{\psi>0,d>0}(w_t, d_t) - \mathcal{P}^w_{\psi>0}(w_t, d_t)$.

In our affine model, we always have $\mathcal{P}^{w}_{\psi>0,d>0}(d_t,w_t)=1$, so changes in intermediary wealth have a linear impact on swap spreads. Again, in our general model, $\mathcal{P}^{w}_{\psi>0,d>0}(d_t,w_t)-\mathcal{P}^{w}_{\psi>0,d<0}(d_t,w_t)\in(-1,1)$ will vary over the state space, so changes in intermediary wealth have a non-linear impact on spreads.

In summary, equations (A-29) and (A-30) show that the non-linearity of our general model stems from the fact that these averaged probabilities—$\mathcal{P}^{d}_{\psi>0}(d_t,w_t)$, $\mathcal{P}^{w}_{\psi>0,d>0}(d_t,w_t)$, $\mathcal{P}^{w}_{\psi>0,d<0}(d_t,w_t)$—as well as $\Delta_{\psi,d=0}(w_{t,},d_t)>0$ vary over the state space. How do these objects vary as a function of w_t and d_t?

First, because the current sign of current demand is informative about the averaged future sign of demand, $\mathcal{P}^{w}_{\psi>0,d>0}(d_t,w_t)-\mathcal{P}^{w}_{\psi>0,d<0}(d_t,w_t)$ is positive (negative) when d_t is sufficiently positive (negative). Specifically, there exists some $d^*(w_t)$, such that $\partial s(d_t,w_t)/\partial w_t>0$ for all $d_t>d^*(w_t)$ and $\partial s(d_t,w_t)/\partial w_t<0$ for all $d_t<d^*(w_t)$. If $\overline{d}>0$, we have $d^*(w_t)<0$. If $\overline{d}<0$, we have $d^*(w_t)>0$. And, if $\overline{d}=0$, we have $d^*(w_t)=0$ and hence $sgn(\partial s(d_t,w_t)/\partial w_t)=sgn(d_t)$.[39]

Second, we have $\partial\mathcal{P}^{d}_{\psi>0}(d_t,w_t)/\partial w_t<0$—increasing the current level of wealth reduces the averaged probability that the constraint binds in the future—and $\partial\Delta_{\psi,d=0}(w_{t,},d_t)/\partial w_t<0$. This in turn implies that

$$\frac{\partial^2 s(d_t,w_t)}{\partial d_t\partial w_t}=-\left(\frac{\kappa_x}{\kappa_o}\right)^2\lambda_o\frac{1-\delta}{1-\delta\rho_d}\cdot\frac{\partial\mathcal{P}^{d}_{\psi>0}(d_t,w_t)}{\partial w_t}-2\kappa_x\frac{1-\delta}{1-\delta\rho_d}\frac{\partial\Delta_{\psi,d=0}(w_{t,},d_t)}{\partial w_t}>0, \tag{A-31}$$

so higher levels of intermediary wealth mute the negative impact of end-user demand on swap spreads.[40]

Third, although both $\partial\mathcal{P}^{w}_{\psi>0,d>0}(d_t,w_t)/\partial w_t<0$ and $\partial\mathcal{P}^{w}_{\psi>0,d<0}(d_t,w_t)/\partial w_t<0$, the first term is larger (smaller) in magnitude when d_t is sufficiently positive (negative). Thus, since

$$\frac{\partial^2 s(d_t,w_t)}{\partial(w_t)^2}=\frac{1}{\kappa_x}\left(\frac{\kappa_x}{\kappa_o}\right)^2\lambda_o\frac{1-\delta}{1-\delta\rho_w}\cdot\left\{\frac{\partial\mathcal{P}^{w}_{\psi>0,d>0}(d_t,w_t)}{\partial w_t}-\frac{\partial\mathcal{P}^{w}_{\psi>0,d<0}(d_t,w_t)}{\partial w_t}\right\}, \tag{A-32}$$

we have $\partial^2 s(d_t,w_t)/\partial(w_t)^2<0$ for d_t sufficiently positive and $\partial^2 s(d_t,w_t)/\partial(w_t)^2>0$ for d_t sufficiently negative.

Finally, since $\mathbb{I}\{\psi_{t+k}>0\}=\mathbb{I}\{\kappa_o\overline{r}_o/\lambda_o+\kappa_x|d_{t+k}|>w_{t+k}\}$—i.e., the constraint depends on the absolute value of demand, we have $\partial\mathcal{P}^{d}_{\psi>0}(d_t,w_t)/\partial d_t>0$ for d_t sufficiently positive and $\partial\mathcal{P}^{d}_{\psi>0}(d_t,w_t)/\partial d_t<0$ for d_t sufficiently negative. Since

$$\frac{\partial^2 s(d_t,w_t)}{\partial(d_t)^2}=-\left(\frac{\kappa_x}{\kappa_o}\right)^2\lambda_o\frac{1-\delta}{1-\delta\rho_d}\cdot\frac{\partial\mathcal{P}^{d}_{\psi>0}(d_t,w_t)}{\partial d_t}-2\kappa_x\frac{1-\delta}{1-\delta\rho_d}\frac{\partial\Delta_{\psi,d=0}(w_{t,},d_t)}{\partial d_t}, \tag{A-33}$$

we obtain $\partial^2 s(d_t,w_t)/\partial(d_t)^2<0$ for d_t sufficiently positive and $\partial^2 s(d_t,w_t)/\partial(d_t)^2>0$ for d_t sufficiently negative.

These comparative statics for $s(d_t,w_t)$ allow us to derive comparative statics for $V(d_t,w_t)=(\delta/(1-\delta))^2\cdot\text{Var}_t[s(d_{t+1},w_{t+1})]$. Lemma 2 (see below) and the foregoing analysis together imply that

$$\frac{\partial V(d_t,w_t)}{\partial d_t}=\left(\frac{\delta}{1-\delta}\right)^2 2\rho_d\text{Cov}_t\left[s(d_{t+1},w_{t+1}),\frac{\partial s(d_{t+1},w_{t+1})}{\partial d_{t+1}}\right] \tag{A-34}$$

will be positive for sufficiently positive values of d_t (since then shocks to both d_{t+1} and w_{t+1} will tend to move $s(d_{t+1},w_{t+1})$ and $\partial s(d_{t+1},w_{t+1})/\partial d_{t+1}$ in the same direction) and will be negative for sufficiently negative values of d_t (then since shocks to both d_{t+1} and w_{t+1} will tend to move $s(d_{t+1},w_{t+1})$ and $\partial s(d_{t+1},w_{t+1})/\partial d_{t+1}$ in the opposite directions). Furthermore, we expect the sign of $\partial V(d_t,w_t)/\partial d_t$ to change at some point $d^{**}(w_t)$ that has the opposite sign as $\overline{d}$—i.e. for a given w_t, $V(d_t,w_t)$ is minimized at $d_t=d^{**}(w_t)$. And, we have $d^{**}(w_t)=0$ when $\overline{d}=0$, so $sgn(\partial V(d_t,w_t)/\partial d_t)=0$, implying that, for a given w_t, $V(d_t,w_t)$ is minimized when $d_t=0$.

[39]For a given w_t, this cutoff function $d^*(w_t)$ is implicity defined by

$$\frac{\mathcal{P}^{w}_{\psi>0,d>0}(d^*(w_t),w_t)}{\mathcal{P}^{w}_{\psi>0}(d^*(w_t),w_t)}=\frac{1}{2},$$

where $\mathcal{P}^{w}_{\psi>0}(d_t,w_t)\equiv(1-\delta\rho_w)\sum_{k=0}^{\infty}(\delta\rho_w)^k\text{Pr}[\psi_{t+k}>0|d_t,w_t]$. Thus, the left-hand-side of the above expression is akin to a conditional probability: it is the (averaged) future probability the constraint binds with positive demand divided by the (averaged) future probability the constraint binds (irrespective of the sign of demand). As a result, we must have $sgn(d^*(w_t))=-sgn(\overline{d})$ for all w_t.

[40]A tedious calculation shows that, consistent with Young's Theorem, we obtain the same expression for $\partial^2 s(w_t,d_t)/\partial d_t\partial w_t$ if we differentiate (A-29) with respect to w_t or if we differentiate (A-30) with respect to d_t.

Furthermore, the comparative statics for $s(d_t, w_t)$ imply that

$$\frac{\partial V(d_t, w_t)}{\partial w_t} = \left(\frac{\delta}{1-\delta}\right)^2 2\rho_w \text{Cov}_t\left[s(d_{t+1}, w_{t+1}), \frac{\partial s(d_{t+1}, w_{t+1})}{\partial w_{t+1}}\right] < 0, \tag{A-35}$$

because shocks to both d_{t+1} and w_{t+1} will tend to move $s(d_{t+1}, w_{t+1})$ and $\partial s(d_{t+1}, w_{t+1})/\partial w_{t+1}$ in the opposite directions. Thus, intuitively, $V(d_t, w_t)$ is greatest when intermediary wealth is low and balance-sheet constraints are more likely to bind in the future.

We summarize these results in the following proposition:

Proposition 9 ***Risk-neutral limit of the non-linear model.*** *Assume that $\alpha = 0$ so intermediaries are risk-neutral, but are subject to potentially binding balance-sheet constraints ($\kappa_x > 0$). Also assume that $\kappa_x \cdot \max\left(\overline{d} + \overline{z}^d, \overline{z}^d - \overline{d}\right) < \overline{w} - \overline{z}^w$, so intermediaries always have sufficient capital to accommodate the exogenous level of end-user demand. However, we assume it is either the case that (i) the sign of end-user demand can fluctuate over time (i.e., we can have $\overline{d} - \overline{z}^d < 0 < \overline{d} + \overline{z}^d$), (ii) intermediaries' balance sheet constraint can be slack (i.e., $\overline{w} + \overline{z}^w > \kappa_o \overline{r}_o/\lambda_o + \kappa_x \min_{z^d \in [-\overline{z}^d, \overline{z}^d]} \left|\overline{d} + z^d\right|$, or both (i) and (ii). Then $s(d_t, w_t)$, $E_t\left[r^s_{t+1}\right]$, and $V(d_t, w_t)$ are non-linear functions that satisfy*

$$\begin{aligned}
\frac{\partial s(d_t, w_t)}{\partial d_t} &< 0, \\
\frac{\partial E_t\left[r^s_{t+1}\right]}{\partial d_t} &< 0 \textit{ and } sgn\left(\frac{\partial E_t[r^s_{t+1}]}{\partial w_t}\right) = sgn(d_t), \\
\frac{\partial V(d_t, w_t)}{\partial w_t} &< 0.
\end{aligned}$$

In addition:

- *When demand always has the same sign, we have $\partial s(d_t, w_t)/\partial w_t = sgn(d_t)$ and $\partial V(d_t, w_t)/\partial d_t = sgn(d_t)$.*
- *When demand can switch sign, there exists some $d^*(w_t)$, whose sign is opposite that of $\overline{d}$, such that we have $\partial s(d_t, w_t)/\partial w_t > 0$ for all $d_t > d^*(w_t)$ and $\partial s(d_t, w_t)/\partial w_t < 0$ for all $d_t < d^*(w_t)$. Similarly, $V(d_t, w_t)$ is a potentially U-shaped function of d_t that achieves its minimum at some value $d^{**}(w_t)$ whose sign is opposite that of $\overline{d}$—i.e., we have $\partial V(d_t, w_t)/\partial d_t > 0$ for $d_t > d^{**}(w_t)$ and $\partial V(d_t, w_t)/\partial d_t < 0$ for $d_t < d^{**}(w_t)$. However, these cutoff thresholds may lie outside of the state space—e.g., when demand is almost always positive, implying that the comparative statics are monotone throughout the state space.*

The general case where $\alpha > 0$. We now consider the general case where $\alpha > 0$ and swap spreads reflect both balance sheet costs and a risk premium. For simplicity, we now revert to our baseline formulation of the constraint. The equilibrium expected return on the swap spread trade is given by

$$\text{E}_t[r^s_{t+1}] = -\kappa_x \cdot sgn(d_t)\,\psi(|d_t|, w_t) - \alpha \cdot V(d_t, w_t)d_t. \tag{A-36}$$

Since $V(d_t, w_t) > 0$ and $\psi(|d_t|, w_t) \geq 0$, we have

$$sgn(\text{E}_t[r^s_{t+1}]) = -sgn(d_t). \tag{A-37}$$

This generalizes our affine model where $sgn(d_t) = 1$ and $\text{E}_t[r^s_{t+1}] < 0$.

How do expected returns vary over the state space? Starting with end-user demand, we write

$$\frac{\partial \text{E}_t[r^s_{t+1}]}{\partial d_t} = \overbrace{(-\kappa_x)\frac{\partial \psi(|d_t|, w_t)}{\partial |d_t|}}^{-\left(\frac{\kappa_x}{\kappa_o}\right)^2 \alpha\sigma_o^2 \cdot \mathbb{I}_{(\psi(|d_t|, w_t) > 0)} \leq 0} + \overbrace{(-\alpha)V(d_t, w_t)}^{<0} + \overbrace{(-\alpha)\frac{\partial V(d_t, w_t)}{\partial d_t}d_t}^{??}.$$

This expression omits the potential discrete jump down in $-\kappa_x sgn(d_t) \cdot \psi(w_t, |d_t|)$ that occurs as d_t crosses from negative to positive. Furthermore, the derivatives of $\psi(|d_t|, w_t)$ with respect to w_t and d_t do not exist on the measure zero set where the leverage constraint binds exactly (i.e., where $\kappa_o \overline{r}_o/\left(\alpha\sigma_o^2\right) + \kappa_x |d_t| - w_t = 0$). The first term is non-positive and the second is negative. However, the sign of the third term is potentially ambiguous when the model is asymmetric (i.e., when $\overline{d} \neq 0$) and end-user demand may switch sign over time (i.e., when $\overline{d} - \overline{z}^d < 0 < \overline{d} + \overline{z}^d$).

Turning to intermediary wealth, we have

$$\frac{\partial \mathrm{E}_t[r^s_{t+1}]}{\partial w_t} = sgn\left(d_t\right) \cdot \overbrace{\left(-\kappa_x \frac{\partial \psi\left(\left|d_t\right|, w_t\right)}{\partial w_t}\right)}^{\geq 0} + d_t \cdot \overbrace{\left(-\alpha \frac{\partial V\left(d_t, w_t\right)}{\partial w_t}\right)}^{>0},$$

which shows that $sgn\left(\partial \mathrm{E}_t[r^s_{t+1}]/\partial w_t\right) = sgn\left(d_t\right)$. Since $sgn(\mathrm{E}_t[r^s_{t+1}]) = -sgn\left(d_t\right)$, this means that increases in intermediary wealth always reduce the absolute value of expected returns—i.e., $\partial\left|\mathrm{E}_t[r^s_{t+1}]\right|/\partial w_t < 0$.

We now turn to the level of spreads. Generalizing equation (A-29) to the case where $\alpha > 0$, we obtain

$$\begin{aligned}\frac{\partial s\left(d_t, w_t\right)}{\partial d_t} = &-\left(\frac{\kappa_x}{\kappa_o}\right)^2 \alpha\sigma_o^2 \frac{1-\delta}{1-\rho_d\delta} \cdot \overbrace{\left\{\left(1-\rho_d\delta\right)\sum\nolimits_{k=0}^{\infty}\left(\delta\rho_d\right)\Pr\left[\psi_{t+k}>0|d_t,w_t\right]\right\}}^{\mathcal{P}^d_{\psi>0}(d_t,w_t)\in(0,1)} \\ &-2\kappa_x\frac{1-\delta}{1-\delta\rho_d}\cdot\underbrace{\left(1-\delta\rho_d\right)\sum\nolimits_{k=0}^{\infty}\left(\delta\rho_d\right)^k \mathrm{E}_t\left[\Delta\left(d_{t+k}\right)\psi(\left|d_{t+k}\right|, w_{t+k})\right]}_{\Delta_{\psi,d=0}(w_t,,d_t)\geq 0} \\ &-\alpha\frac{1-\delta}{1-\rho_d\delta}\cdot\left(1-\rho_d\delta\right)\sum\nolimits_{k=0}^{\infty}\left(\delta\rho_d\right)^k \mathrm{E}_t\left[\underbrace{V(d_{t+k},w_{t+k})}_{>0}+\underbrace{\frac{\partial V(d_{t+k},w_{t+k})}{\partial d_{t+k}}d_{t+k}}_{??}\right].\end{aligned}$$

There are fours term in the above expression. The signs of the first three are unambiguous in our general model. However, the fourth term, which involves $\left[\partial V(d_{t+k}, w_{t+k})/\partial d_{t+k}\right] d_{t+k}$, is potentially ambiguous when the model is asymmetric (i.e., when $\overline{d} \neq 0$) and end-user demand may switch sign over time (i.e., when $\overline{d} - \overline{z}^d < 0 < \overline{d} + \overline{z}^d$).[41]

Turning to the impact of wealth, a similar calculation shows that:

$$\begin{aligned}\frac{\partial s\left(d_t, w_t\right)}{\partial w_t} = &\frac{1}{\kappa_x}\left(\frac{\kappa_x}{\kappa_o}\right)^2 \alpha\sigma_o^2\frac{1-\delta}{1-\rho_w\delta}\cdot\overbrace{\left\{\left(1-\delta\rho_w\right)\sum\nolimits_{k=0}^{\infty}\left(\delta\rho_w\right)^k\left(\begin{array}{c}\Pr\left[\psi_{t+k}>0,d_{t+k}>0|d_t,w_t\right]\\ -\Pr\left[\psi_{t+k}>0,d_{t+k}<0|d_t,w_t\right]\end{array}\right)\right\}}^{\mathcal{P}^w_{\psi>0,d>0}(d_t,w_t)-\mathcal{P}^w_{\psi>0,d<0}(d_t,w_t)\in(-1,1)} \\ &-\alpha\frac{1-\delta}{1-\delta\rho_w}\cdot\left(1-\delta\rho_w\right)\sum\nolimits_{k=0}^{\infty}\left(\delta\rho_w\right)^k\mathrm{E}_t\left[\underbrace{\frac{\partial V(d_{t+k},w_{t+k})}{\partial w_{t+k}}}_{<0}d_{t+k}\right].\end{aligned} \quad \text{(A-38)}$$

Equation (A-38) shows that the terms for time $t+k$ depend on the likely sign of end-user demand at d_{t+k}.[42] Thus, as above, there should exist some $d^*\left(w_t\right)$, such that $\partial s\left(d_t, w_t\right)/\partial w_t > 0$ for all $d_t > d^*\left(w_t\right)$ and $\partial s\left(d_t, w_t\right)/\partial w_t < 0$ for all $d_t < d^*\left(w_t\right)$. If $\overline{d} > 0$, we have $d^*\left(w_t\right) < 0$. If $\overline{d} < 0$, we have $d^*\left(w_t\right) > 0$. And, if $\overline{d} = 0$, we have $d^*\left(w_t\right) = 0$ and hence $sgn\left(\partial s\left(d_t, w_t\right)/\partial w_t\right) = sgn\left(d_t\right)$.

Special cases of the non-linear model. To make more precise predictions about $\partial s_t/\partial d_t$ and $\partial s_t/\partial w_t$, we need to make further assumptions. In particular, we need to rule out cases where demand is expected to switch sign in the near term.

As summarized in the following proposition, one sufficient condition is to assume that, while the constraint may or may not bind in the future, the sign of d_{t+k} is always positive (or negative). This yields $\partial s_t/\partial d_t < 0$ and $sgn(\partial s_t/\partial w_t) = sgn(d_t)$, in line with the results of our linear model.

Proposition 10 ***An asymmetric, non-linear special case.*** *Assume (i) $0 < \overline{d} - \overline{z}^d$ so end-user demand to receive fixed is always positive, (ii) $\kappa_x\left(\overline{d} + \overline{z}^d\right) < \overline{w} - \overline{z}^w$ so intermediaries always have enough capital to accommodate end-user demand, but (iii) $\overline{w} + \overline{z}^w > \kappa_o\overline{r}_o/\left(\alpha\sigma_o^2\right) + \kappa_x\left(\overline{d} - \overline{z}^d\right)$ so intermediaries' constraint is slack when intermediary wealth is large relative to the end-user demand. Then $s\left(d_t, w_t\right)$, $\mathrm{E}_t\left[r^s_{t+1}\right]$, and $V\left(d_t, w_t\right)$ are non-linear functions of*

[41]When $\mathcal{P}^d_{\psi>0}\left(d_t, w_t\right) = 1$ and $\Delta_{\psi,d=0}\left(w_t,, d_t\right) = 0$, this reduces to our affine model. Specifically, the second and fourth terms are zero, the first term is $-\left(\left(1-\delta\right)/\left(1-\rho_d\delta\right)\right)\alpha\left(\kappa_x/\kappa_o\right)^2\sigma_o^2 < 0$, and the third term is $-\left(\left(1-\delta\right)/\left(1-\rho_d\delta\right)\right)\alpha V < 0$.

[42]When $\mathcal{P}^w_{\psi>0,d>0}\left(d_t, w_t\right) = 1$, this reduces to our affine model. Specifically, the first term reduces to $\left(\left(1-\delta\right)/\left(1-\rho_w\delta\right)\right)\frac{1}{\kappa_x}\alpha\left(\kappa_x/\kappa_o\right)^2\sigma_o^2 > 0$ and the second term vanishes since then $\partial V(w_t, d_t)/\partial w_t = 0$.

the state variables that satisfy

$$\begin{aligned}\frac{\partial s(d_t,w_t)}{\partial d_t} &< 0 \text{ and } \frac{\partial s(d_t,w_t)}{\partial w_t} > 0,\\ \frac{\partial E_t\left[r^s_{t+1}\right]}{\partial d_t} &< 0 \text{ and } \frac{\partial E\left[r^s_{t+1}\right]}{\partial w_t} > 0,\\ \frac{\partial V(d_t,w_t)}{\partial d_t} &> 0 \text{ and } \frac{\partial V(d_t,w_t)}{\partial w_t} < 0.\end{aligned}$$

Alternately, we can also allow demand to switch sign, but make the model symmetric by assuming $\overline{d}$ =E$[d_t] = 0$. In this case, the future sign of d_{t+k} is more likely than not to be the same the current sign of d_t—i.e., we have Pr$[sgn\,(d_{t+k}) = sgn\,(d_t)\,|d_t, w_t] > 1/2$ unless $d_t = 0$—and we again have $\partial s_t/\partial d_t < 0$ and $sgn(\partial s_t/\partial w_t) = sgn(d_t)$.

Proposition 11 ***A symmetric, non-linear special case.*** *Assume (i) the steady-state level of end-user demand is zero (i.e.,* $\overline{d} = 0$*) so the model is symmetric, (ii)* $\kappa_x\overline{z}^d < \overline{w} - \overline{z}^w$ *so intermediaries always have enough capital to accommodate end-user demand, but (iii)* $\overline{w} + \overline{z}^w > \kappa_o\overline{r}_o/\alpha\sigma_o^2$ *so intermediaries constraint can be slack when wealth is large and the magnitude of end-user demand* $|d_t|$ *is small. Then* $s\,(d_t, w_t)$*,* $E_t\left[r^s_{t+1}\right]$*, and* $V(d_t, w_t)$ *are non-linear functions of the state variables that satisfy:*

$$\begin{aligned}\frac{\partial s(d_t,w_t)}{\partial d_t} &< 0 \text{ and } \frac{\partial s(d_t,w_t)}{\partial w_t} = sgn\,(d_t)\,,\\ \frac{\partial E_t\left[r^s_{t+1}\right]}{\partial d_t} &< 0 \text{ and } \frac{\partial E\left[r^s_{t+1}\right]}{\partial w_t} = sgn\,(d_t)\,,\\ \frac{\partial V(d_t,w_t)}{\partial d_t} &= sgn\,(d_t) \text{ and } \frac{\partial V(d_t,w_t)}{\partial w_t} < 0.\end{aligned}$$

Generalizing beyond these special cases. Since the model's stable equilibrium is continuous in the underlying model parameters, we can use the various special cases characterized above to say more about how the model behaves for other parameter values. For instance, if the probability that end-user demand turns negative is sufficiently small, then by continuity the solution must have the same properties as those in *Case 2*. Similarly, if either (i) α is sufficiently small for any value of $\overline{d} \neq 0$ or (ii) if $\left|\overline{d}\right|$ is sufficiently small for any value of $\alpha > 0$, then by continuity the solution must have the same properties as in *Case 4*.[43] And, while the above analysis assumes that $\gamma = 0$, this characterization remains valid so long as the demand elasticity γ is not too large.

Overall, the key qualitative implications of our affine model seem to generalize nicely once we allow for nonlinearities. Indeed, it is clear that the global behavior of $s\,(d_t, w_t)$, E$_t\left[r^s_{t+1}\right]$, and $V(d_t, w_t)$ that we see in these special cases must obtain for any set of model parameters. The only thing that we have been unable to rule out—either analytically or numerically—is that the local behavior of $s\,(d_t, w_t)$, E$_t\left[r^s_{t+1}\right]$, and $V(d_t, w_t)$ may conflict with what we see in the special cases. For instance, we have not been able to rule out the possibility that $s\,(d_t, w_t)$ could be locally increasing in d_t in some case where α is large and $\overline{d}$ is large (but not too large since then demand will rarely be negative and the solution will be like that in *Case 2*). However, even if our theory predicted small local monotonicities of this sort for certain parameterizations, we would be inclined to view this as more of a theoretical curiosity than as an empirical prediction with important practical relevance.[44]

B.3.6 Proofs

We make repeated use of the following result:

Lemma 2 *Suppose the random variable* $s\,(d_{t+1}, w_{t+1})$ *is a differentiable function of* d_{t+1} *and* w_{t+1} *which follow the processes given above. Then we have*

$$\frac{\partial}{\partial d_t}Var_t\left[s\,(d_{t+1}, w_{t+1})\right] = 2\rho_d Cov\left[s\,(d_{t+1}, w_{t+1})\,, \frac{\partial s\,(d_{t+1}, w_{t+1})}{\partial d_{t+1}}\right] \qquad \text{(A-39)}$$

[43]For instance, we have signed the comparative statics for $s\,(d_t, w_t)$ and $V\,(d_t, w_t)$ in limit where $\alpha = 0$. By continuity of the model's stable solution in α, these comparative statics must also have the same sign for small values of α. Similarly, we have signed the comparative statics in the case where $\overline{d} = 0$ and these comparative statics will change continuously as we make $\overline{d}$ slightly positive or slightly negative.

[44]Specifically, any such local non-monotonicities would rely on almost heroic levels of investor foresight and investor knowledge about the underlying data generating processes. And, we believe that treating these as empirical predictions that we should take to the data would be taking our model far too literally.

and

$$\frac{\partial}{\partial w_t} Var_t\left[s\left(d_{t+1}, w_{t+1}\right)\right] = 2\rho_w Cov\left[s\left(d_{t+1}, w_{t+1}\right), \frac{\partial s\left(d_{t+1}, w_{t+1}\right)}{\partial w_{t+1}}\right]. \tag{A-40}$$

Proof. We have

$$\begin{aligned}
\frac{\partial}{\partial d_t}\text{Var}_t\left[s\left(d_{t+1}, w_{t+1}\right)\right] &= \frac{\partial}{\partial d_t}\left(\text{E}_t\left[s^2\left(d_{t+1}, w_{t+1}\right)\right] - \left(\text{E}\left[s\left(d_{t+1}, w_{t+1}\right)\right]\right)^2\right) \\
&= 2\text{E}_t\left[s\left(d_{t+1}, w_{t+1}\right)\frac{\partial s\left(d_{t+1}, w_{t+1}\right)}{\partial d_{t+1}}\frac{\partial d_{t+1}}{\partial d_t}\right] - 2\text{E}_t\left[s\left(d_{t+1}, w_{t+1}\right)\right]\text{E}_t\left[\frac{\partial s\left(d_{t+1}, w_{t+1}\right)}{\partial d_{t+1}}\frac{\partial d_{t+1}}{\partial d_t}\right] \\
&= 2\rho_d\text{Cov}\left[s\left(d_{t+1}, w_{t+1}\right), \frac{\partial s\left(d_{t+1}, w_{t+1}\right)}{\partial d_{t+1}}\right].
\end{aligned}$$

The expression for $\frac{\partial}{\partial w_t}\text{Var}_t\left[s\left(d_{t+1}, w_{t+1}\right)\right]$ follows from an identical argument. ■

Proof of Proposition 11. We consider the following set of recursively-defined functions:

$$s^{(k)}\left(d_t, w_t\right) \equiv (1-\delta)\left[\begin{array}{c}(-\kappa_x) sgn\left(d_t\right)\cdot\psi\left(\left|d_t\right|, w_t\right) \\ +(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2\mathcal{V}^{(k)}(d_t, w_t)\cdot d_t\end{array}\right] + \delta\text{E}\left[s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)|d_t, w_t\right], \tag{A-41}$$

where, for $k \geq 2$, we have

$$\mathcal{V}^{(k)}(d_t, w_t) \equiv \text{Var}\left[s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)|d_t, w_t\right] \tag{A-42}$$

and

$$s^{(1)}\left(d_t, w_t\right) \equiv -\kappa_x sgn\left(d_t\right)\cdot\psi\left(\left|d_t\right|, w_t\right). \tag{A-43}$$

We have

$$s_t = s\left(d_t, w_t\right) = \lim_{k\to\infty} s^{(k)}\left(d_t, w_t\right). \tag{A-44}$$

In other words, the stable equilibrium of our infinite horizon model is just the limit of this recursive set of functions as $k\to\infty$. (Other equilibrium, if they exist, correspond to different starting points $s^{(1)}\left(d_t, w_t\right)$.)

Formally, consider the functional operator $S[\cdot]$:

$$S\left[s\left(d_t, w_t\right)\right] \equiv (1-\delta)\left[\begin{array}{c}(-\kappa_x) sgn\left(d_t\right)\cdot\psi\left(\left|d_t\right|, w_t\right) \\ +(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2\text{Var}\left[s\left(d_{t+1}, w_{t+1}\right)|d_t, w_t\right]\cdot d_t\end{array}\right] + \delta\text{E}\left[s\left(d_{t+1}, w_{t+1}\right)|d_t, w_t\right]. \tag{A-45}$$

This operator takes as its argument a function $s\left(d_t, w_t\right)$ and returns a new function $S\left[s\left(d_t, w_t\right)\right]$.

When $\alpha = 0$, this functional operator satisfies Blackwell's Condition—i.e., monotonicity (if $f(\mathbf{z}) \leq g(\mathbf{z})$ then $S[f(\mathbf{z})] \leq S[g(\mathbf{z})]$) and discounting (for an arbitrary constant c and function $f(\mathbf{z})$, $S[f(\mathbf{z})+c] \leq S[f(\mathbf{z})] + \delta c$ where $\delta < 1$)—and is thus a contraction mapping. As a result, the Contraction Mapping Theorem implies there is a unique solution to $s\left(d_t, w_t\right) = S\left[s\left(d_t, w_t\right)\right]$ which is simply:

$$s\left(d_t, w_t\right) = (1-\delta)\sum_{k=0}^{\infty}\delta^k\text{E}\left[(-\kappa_x) sgn\left(d_{t+k}\right)\psi\left(\left|d_{t+k}\right|, w_{t+k}\right)|d_t, w_t\right]. \tag{A-46}$$

Furthermore, $s\left(d_t, w_t\right) = \lim_{k\to\infty} S^k\left[s^{(1)}\left(d_t, w_t\right)\right]$ for any arbitrary starting function $s^{(1)}\left(d_t, w_t\right)$.

When $\alpha > 0$, this functional operator no longer satisfies Blackwell's Conditions. And, for large values of α, $S[\cdot]$ need not be a contraction mapping, so multiple solutions to $s\left(d_t, w_t\right) = S\left[s\left(d_t, w_t\right)\right]$ may exist. Specifically, even though $\lim_{k\to\infty} S^k\left[s^{(1)}\left(d_t, w_t\right)\right]$ will always exist for an arbitrary starting function $s^{(1)}\left(d_t, w_t\right)$, we may reach different limiting solutions, $s\left(d_t, w_t\right) = \lim_{k\to\infty} S^k\left[s^{(1)}\left(d_t, w_t\right)\right]$, depending on the choice of the initial condition $s^{(1)}\left(d_t, w_t\right)$. However, the stable equilibrium of our model, which always exists, corresponds to $s_t^{(1)} = -\kappa_x sgn\left(d_t\right)\cdot\psi\left(\left|d_t\right|, w_t\right)$.

For all k, we want to show that

$$\mathcal{V}^{(k)}(d_t, w_t) = \mathcal{V}^{(k)}(\left|d_t\right|, w_t) > 0, \tag{A-47a}$$

$$\frac{\partial\mathcal{V}^{(k)}(d_t, w_t)}{\partial w_t} < 0 \text{ and } sgn\left(\frac{\partial\mathcal{V}^{(k)}(d_t, w_t)}{\partial d_t}\right) = sgn\left(d_t\right), \text{ whereas} \tag{A-47b}$$

$$\frac{\partial^2\mathcal{V}^{(k)}(d_t, w_t)}{\partial\left(w_t\right)^2} \geq 0, \frac{\partial^2\mathcal{V}^{(k)}(d_t, w_t)}{\partial\left(d_t\right)^2} \geq 0, \text{ and } sgn\left(\frac{\partial^2\mathcal{V}^{(k)}(d_t, w_t)}{\partial w_t\partial d_t}\right) = -sgn\left(d_t\right). \tag{A-47c}$$

We also want to show that

$$\frac{\partial s^{(k)}(d_t, w_t)}{\partial w_t} = sgn(d_t) \text{ and } \frac{\partial s^{(k)}(d_t, w_t)}{\partial d_t} < 0, \text{ whereas} \tag{A-48a}$$

$$sgn\left(\frac{\partial^2 s^{(k)}(d_t, w_t)}{\partial (d_t)^2}\right) = sgn\left(\frac{\partial^2 s^{(k)}(d_t, w_t)}{\partial (w_t)^2}\right) = -sgn(d_t) \text{ and } \frac{\partial^2 s^{(k)}(d_t, w_t)}{\partial d_t \partial w_t} \geq 0. \tag{A-48b}$$

To prove this inductively we need to:

1. Prove that properties (A-47a), (A-47b), (A-47c), (A-48a), and (A-48b) hold for $k = 2$.
2. Prove that if all of these properties hold for $k-1$, then they also hold for k.

Analysis of 1-period swap spreads. We have

$$s_t^{(1)} = -\kappa_x sgn(d_t) \cdot \psi(|d_t|, w_t)$$

where

$$\psi(|d_t|, w_t) = \max\left\{0, \frac{\alpha\sigma_o^2}{\kappa_o^2}\left(\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x |d_t| - w_t\right)\right\} \geq 0.$$

Note that

$$\begin{aligned} \frac{\partial \psi(|d_t|, w_t)}{\partial d_t} &= \frac{\alpha\sigma_o^2}{\kappa_o^2}\kappa_x sgn(d_t)\,\mathbb{I}\left\{\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x |d_t| - w_t > 0\right\} \text{ and} \\ \frac{\partial \psi(|d_t|, w_t)}{\partial w_t} &= -\frac{\alpha\sigma_o^2}{\kappa_o^2}\mathbb{I}\left\{\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x |d_t| - w_t > 0\right\} \leq 0, \end{aligned}$$

whereas

$$\frac{\partial^2 \psi(|d_t|, w_t)}{\partial (d_t)^2} \overset{\text{a.s.}}{=} \frac{\partial^2 \psi(|d_t|, w_t)}{\partial d_t \partial w_t} \overset{\text{a.s.}}{=} \frac{\partial^2 \psi(|d_t|, w_t)}{\partial (w_t)^2} \overset{\text{a.s.}}{=} 0.$$

Thus, we have[45]

$$\frac{\partial s_t^{(1)}}{\partial d_t} = -\kappa_x sgn(d_t) \cdot \frac{\partial \psi(|d_t|, w_t)}{\partial d_t} = -\kappa_x^2 [sgn(d_t)]^2 \cdot \frac{\alpha\sigma_o^2}{\kappa_o^2}\mathbb{I}\left\{\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x |d_t| - w_t > 0\right\} \leq 0$$

and

$$\frac{\partial s_t^{(1)}}{\partial w_t} = -\kappa_x sgn(d_t) \cdot \frac{\partial \psi(|d_t|, w_t)}{\partial w_t} = sgn(d_t) \cdot \kappa_x \frac{\alpha\sigma_o^2}{\kappa_o^2}\mathbb{I}\left\{\kappa_o \frac{\overline{r}_o}{\alpha\sigma_o^2} + \kappa_x |d_t| - w_t > 0\right\}.$$

We conclude that

$$\frac{\partial s^{(1)}(d_t, w_t)}{\partial d_t} \leq 0 \text{ and } \frac{\partial s^{(1)}(d_t, w_t)}{\partial w_t} = sgn(d_t),$$

while

$$\frac{\partial^2 s^{(1)}(d_t, w_t)}{\partial (d_t)^2} \overset{\text{a.s.}}{=} \frac{\partial^2 s^{(1)}(d_t, w_t)}{\partial d_t \partial w_t} \overset{\text{a.s.}}{=} \frac{\partial^2 s^{(1)}(d_t, w_t)}{\partial (w_t)^2} \overset{\text{a.s.}}{=} 0.$$

Analysis of 2-period swap spreads.

Characterizing $V^{(2)}(d_t, w_t)$: Since

$$\mathcal{V}^{(2)}(d_t, w_t) = \text{Var}\left[s^{(1)}(d_{t+1}, w_{t+1})\,|d_t, w_t\right] \overset{a.s.}{=} (\kappa_x)^2\,\text{Var}\left[\psi(|d_{t+1}|, w_{t+1})\,|d_t, w_t\right] \geq 0,$$

Lemma 2 implies

$$\frac{\partial \mathcal{V}^{(2)}(d_t, w_t)}{\partial w_t} = 2\rho_w \kappa_x^2 \text{Cov}_t\left[\psi(|d_{t+1}|, w_{t+1}), \frac{\partial \psi(|d_{t+1}|, w_{t+1})}{\partial w_{t+1}}\right].$$

As of time t, the only random variables determining $\psi(|d_{t+1}|, w_{t+1})$ and $\partial\psi(|d_{t+1}|, w_{t+1})/\partial w_{t+1}$ are $\varepsilon_{w,t+1}$ and $\varepsilon_{d,t+1}$. In particular, $\psi(|d_{t+1}|, w_{t+1})$ is weakly decreasing in w_{t+1} and $\partial\psi(|d_{t+1}|, w_{t+1})/\partial w_{t+1}$ is weakly increasing in w_{t+1}. Similarly, $\psi(|d_{t+1}|, w_{t+1})$ is weakly increasing (decreasing) in d_{t+1} and $\partial\psi(|d_{t+1}|, w_{t+1})/\partial w_{t+1}$ is

[45] This expression omits the potential discrete jump down in $-\kappa_x sgn(d_t) \cdot \psi(w_t, |d_t|)$ that occurs as d_t crosses from negative to positive.

weakly decreasing (increasing) in w_{t+1} when d_{t+1} is positive (negative). Thus, shocks to both w_{t+1} and d_{t+1} have opposing effects on $\psi(|d_{t+1}|, w_{t+1})$ and $\partial\psi(|d_{t+1}|, w_{t+1})/\partial w_{t+1}$, so we have

$$\frac{\partial \mathcal{V}^{(2)}(d_t, w_t)}{\partial w_t} = 2\rho_w \kappa_x^2 \text{Cov}_t \left[\psi(|d_{t+1}|, w_{t+1}), \frac{\partial\psi(|d_{t+1}|, w_{t+1})}{\partial w_{t+1}} \right] < 0. \quad \text{(A-49)}$$

Technically, when k is small (e.g., for $k = 2$), we cannot rule out the possibility that $V^{(k)}(d_t, w_t) = 0$ at some points in the state space.[46] This is because the shocks to d_{t+1} and w_{t+1} have bounded support, so there could be points in the state-space where $\Pr_t[\psi(|d_{t+1}|, w_{t+1}) = 0] = 1$. However, we cannot have $\mathcal{V}^{(k)}(d_t, w_t) = 0$ for k large because it is possible to travel from any interior point in the state space to any other interior point in a finite number of steps. Similarly, it is possible to travel from any point on the boundary of the state space to any point in the interior in a finite number of steps. Thus, for k large, there will always be news about the future path of balance sheet costs and hence we will have $V^{(k)}(d_t, w_t) > 0$ everywhere in the state space. For simplicity, we ignore these issues in what follows and state inequalities that are potentially weak for k small (but necessarily strict for k large) as being strict for all k.

Using a similar argument to the one we used for $\partial\mathcal{V}^{(2)}(d_t, w_t)/\partial w_t$, we have

$$\frac{\partial \mathcal{V}^{(2)}(d_t, w_t)}{\partial d_t} = 2\rho_d \kappa_x^2 \text{Cov}_t \left[\psi(|d_{t+1}|, w_{t+1}), \frac{\partial\psi(|d_{t+1}|, w_{t+1})}{\partial d_{t+1}} \right].$$

The only random variables determining $\psi(|d_{t+1}|, w_{t+1})$ and $\partial\psi(|d_{t+1}|, w_{t+1})/\partial d_{t+1}$ are $\varepsilon_{w,t+1}$ and $\varepsilon_{d,t+1}$. In particular, $\psi(|d_{t+1}|, w_{t+1})$ is always weakly decreasing in w_{t+1} and $\partial\psi(|d_{t+1}|, w_{t+1})/\partial d_{t+1}$ is weakly decreasing (increasing) in w_{t+1} when d_{t+1} is positive (negative). Similarly, $\psi(|d_{t+1}|, w_{t+1})$ is increasing (decreasing) in d_{t+1} when d_{t+1} is positive (negative) and $\partial\psi(|d_{t+1}|, w_{t+1})/\partial d_{t+1}$ is always weakly increasing in d_{t+1}. Therefore, shocks to both d_{t+1} and w_{t+1} lead $\psi(|d_{t+1}|, w_{t+1})$ and $\partial\psi(|d_{t+1}|, w_{t+1})/\partial d_{t+1}$ to comove positively when $d_{t+1} > 0$. By contrast, these shocks lead $\psi(|d_{t+1}|, w_{t+1})$ and $\partial\psi(|d_{t+1}|, w_{t+1})/\partial d_{t+1}$ to comove negatively when $d_{t+1} < 0$. Because $\overline{d} = 0$, it is likely that d_{t+1} and d_t have the same sign (i.e., $\Pr_t[sgn(d_{t+1}) = sgn(d_t)] > 1/2$ unless $d_t = 0$), so we have

$$sgn\left(\frac{\partial \mathcal{V}^{(2)}(d_t, w_t)}{\partial d_t}\right) = sgn\left(\text{Cov}_t\left[\psi(|d_{t+1}|, w_{t+1}), \frac{\partial\psi(|d_{t+1}|, w_{t+1})}{\partial d_{t+1}}\right]\right) = sgn(d_t). \quad \text{(A-50)}$$

Furthermore, by the symmetry of the d_t process, we have $\mathcal{V}^{(2)}(d_t, w_t) = \mathcal{V}^{(2)}(w_t, -d_t) = \mathcal{V}^{(2)}(|d_t|, w_t)$.

We also have

$$\frac{\partial^2 \mathcal{V}^{(2)}(d_t, w_t)}{\partial (d_t)^2} = 2\rho_d^2 \kappa_x^2 \text{Var}_t\left[\frac{\partial\psi(|d_{t+1}|, w_{t+1})}{\partial d_{t+1}}\right] > 0, \quad \text{(A-51)}$$

$$\frac{\partial^2 \mathcal{V}^{(2)}(d_t, w_t)}{\partial (w_t)^2} = 2\rho_w^2 \kappa_x^2 \text{Var}_t\left[\frac{\partial\psi(|d_{t+1}|, w_{t+1})}{\partial w_{t+1}}\right] > 0, \quad \text{(A-52)}$$

$$\begin{aligned}\frac{\partial^2 \mathcal{V}^{(2)}(d_t, w_t)}{\partial d_t \partial w_t} &= 2\rho_d\rho_w \kappa_x^2 \text{Cov}_t\left[\frac{\partial\psi(|d_{t+1}|, w_{t+1})}{\partial w_{t+1}}, \frac{\partial\psi(|d_{t+1}|, w_{t+1})}{\partial d_{t+1}}\right] \quad \text{(A-53)} \\ &= -2\rho_d\rho_w\kappa_x^2\left(\frac{\alpha\sigma_o^2}{\kappa_o^2}\right)\left(\frac{\alpha\sigma_o^2}{\kappa_o^2}\kappa_x\right)\text{Cov}_t\left[\mathbb{I}\{\psi_{t+1} > 0\}, sgn(d_{t+1})\mathbb{I}\{\psi_{t+1} > 0\}\right]\end{aligned}$$

Thus, as $sgn(d_t)$ is informative about $sgn(d_{t+1})$, we have $sgn\left(\partial^2\mathcal{V}^{(2)}(d_t, w_t)/\partial w_t \partial d_t\right) = -sgn(d_t)$.

Characterizing $s^{(2)}(d_t, w_t)$: After some algebra, we have[47]

$$\begin{aligned}\frac{\partial s^{(2)}(d_t, w_t)}{\partial d_t} &= (1-\delta)\left[\begin{array}{c}(-\kappa_x)[sgn(d_t)]^2 \cdot \frac{\partial\psi(|d_t|, w_t)}{\partial|d_t|} \\ +(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2 \mathcal{V}^{(2)}(d_t, w_t) + (-\alpha)\left(\frac{\delta}{1-\delta}\right)^2 \frac{\partial\mathcal{V}^{(2)}(d_t, w_t)}{\partial d_t} d_t\end{array}\right] \quad \text{(A-54)} \\ &+ \delta\rho_d \text{E}_t\left[\frac{\partial s^{(1)}(d_{t+1}, w_{t+1})}{\partial d_{t+1}}\right] < 0,\end{aligned}$$

[46]Of course, we could immediately rule this out if we reintroduced the stochastic fundamental component of swap spreads, m_t.

[47]This expression omits the potential discrete jump down in $-\kappa_x sgn(d_t) \cdot \psi(w_t, |d_t|)$ that occurs as d_t crosses from negative to positive.

whereas

$$\frac{\partial s^{(2)}\left(d_t,w_t\right)}{\partial w_t}=(1-\delta)\left[\begin{array}{c}(-\kappa_x)sgn\left(d_t\right)\cdot\frac{\partial\psi(|d_t|,w_t)}{\partial w_t}\\+(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2\frac{\partial V^{(2)}(d_t,w_t)}{\partial w_t}\cdot d_t\end{array}\right]+\delta\rho_w\mathrm{E}_t\left[\frac{\partial s^{(1)}\left(d_{t+1},w_{t+1}\right)}{\partial w_{t+1}}\right]. \tag{A-55}$$

The sign of the first term in square brackets is the same as that of d_t. Moreover, we have $sgn\left(\partial s^{(1)}\left(d_{t+1},w_{t+1}\right)/\partial w_{t+1}\right)=sgn\left(d_{t+1}\right)$ and, since $\Pr_t\left[sgn\left(d_{t+1}\right)=sgn\left(d_t\right)\right]>1/2$ unless $d_t=0$, we have $sgn\left(\mathrm{E}_t\left[\partial s^{(1)}\left(d_{t+1},w_{t+1}\right)/\partial\partial w_{t+1}\right]\right)=sgn\left(d_t\right)$. Therefore, $sgn\left(\partial s^{(2)}\left(d_t,w_t\right)/\partial w_t\right)=sgn\left(d_t\right)$.

Next note that

$$\frac{\partial^2 s^{(2)}\left(d_t,w_t\right)}{\partial\left(d_t\right)^2}\overset{\text{a.s.}}{=}(1-\delta)\left[2(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2\frac{\partial\mathcal{V}^{(2)}(d_t,w_t}{\partial d_t}+(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2\frac{\partial^2\mathcal{V}^{(2)}(d_t,w_t)}{\partial\left(d_t\right)^2}d_t\right], \tag{A-56}$$

so $sgn(\partial^2 s^{(2)}\left(d_t,w_t\right)/\partial\left(d_t\right)^2)=-sgn\left(d_t\right)$. Similarly, we have

$$\frac{\partial^2 s^{(2)}\left(d_t,w_t\right)}{\partial\left(w_t\right)^2}=(1-\delta)\left[(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2\frac{\partial^2 V^{(2)}(d_t,w_t)}{\partial\left(w_t\right)^2}\cdot d_t\right], \tag{A-57}$$

so $sgn(\partial^2 s^{(2)}\left(d_t,w_t\right)/\partial\left(w_t\right)^2)=-sgn\left(d_t\right)$. Finally, we have

$$\frac{\partial^2 s^{(2)}\left(d_t,w_t\right)}{\partial w_t\partial d_t}=(1-\delta)\left[(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2\left(\frac{\partial V^{(2)}(d_t,w_t)}{\partial w_t}+\frac{\partial^2 V^{(2)}(d_t,w_t)}{\partial w_t\partial d_t}\cdot d_t\right)\right]>0. \tag{A-58}$$

Thus, we have proved that properties (A-47a), (A-47b), (A-47c), (A-48a), and (A-48b) hold for $k=2$.

Inductive step: Analysis of k-period swap spreads.

We now show that if properties (A-47a), (A-47b), (A-47c), (A-48a), and (A-48b) hold for $k-1$, then they also hold for k.

Characterizing $V^{(k)}\left(d_t,w_t\right)$: We have

$$\frac{\partial\mathcal{V}^{(k)}(d_t,w_t)}{\partial d_t}=2\rho_d\mathrm{Cov}_t\left[s^{(k-1)}\left(d_{t+1},w_{t+1}\right),\frac{\partial s^{(k-1)}\left(d_{t+1},w_{t+1}\right)}{\partial d_{t+1}}\right].$$

By our inductive hypothesis, we have

$$\frac{\partial s^{(k-1)}\left(d_{t+1},w_{t+1}\right)}{\partial d_{t+1}}<0\text{ and }sgn\left(\frac{\partial^2 s^{(k-1)}\left(d_{t+1},w_{t+1}\right)}{\partial\left(d_{t+1}\right)^2}\right)=-sgn\left(d_{t+1}\right)\text{, while}$$
$$\frac{\partial s^{(k-1)}\left(d_{t+1},w_{t+1}\right)}{\partial w_{t+1}}=sgn\left(d_{t+1}\right)\ \text{ and }\ \frac{\partial^2 s^{(k-1)}\left(d_{t+1},w_{t+1}\right)}{\partial d_{t+1}\partial w_{t+1}}>0,$$

so shocks to both d_{t+1} and w_{t+1} lead $s^{(k-1)}\left(d_{t+1},w_{t+1}\right)$ and $\partial s^{(k-1)}\left(d_{t+1},w_{t+1}\right)/\partial d_{t+1}$ to comove positively when $d_{t+1}>0$. By contrast, these shocks lead $s^{(k-1)}\left(d_{t+1},w_{t+1}\right)$ and $\partial s^{(k-1)}\left(d_{t+1},w_{t+1}\right)/\partial d_{t+1}$ to comove negatively when $d_{t+1}<0$. Because $\overline{d}=0$, it is likely that d_{t+1} and d_t have the same sign (i.e., $\Pr_t\left[sgn\left(d_{t+1}\right)=sgn\left(d_t\right)\right]>1/2$ unless $d_t=0$), so we have

$$sgn\left(\frac{\partial V^{(k)}(d_t,w_t)}{\partial d_t}\right)=sgn\left(d_t\right). \tag{A-59}$$

Next we have

$$\frac{\partial\mathcal{V}^{(k)}(d_t,w_t)}{\partial w_t}=2\rho_w\mathrm{Cov}_t\left[s^{(k-1)}\left(d_{t+1},w_{t+1}\right),\frac{\partial s^{(k-1)}\left(d_{t+1},w_{t+1}\right)}{\partial w_{t+1}}\right].$$

By our inductive hypothesis, we have

$$\frac{\partial s^{(k-1)}\left(d_{t+1},w_{t+1}\right)}{\partial d_{t+1}}<0\text{ and }\frac{\partial^2 s^{(k-1)}\left(d_{t+1},w_{t+1}\right)}{\partial d_{t+1}\partial w_{t+1}}>0\text{, while}$$
$$\frac{\partial s^{(k-1)}\left(d_{t+1},w_{t+1}\right)}{\partial w_{t+1}}=sgn\left(d_{t+1}\right)\ \text{ and }sgn\left(\frac{\partial^2 s^{(k-1)}\left(d_{t+1},w_{t+1}\right)}{\partial\left(w_{t+1}\right)^2}\right)=-sgn\left(d_{t+1}\right),$$

so shocks to both d_{t+1} and w_{t+1} lead $s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)$ and $\partial s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)/\partial w_{t+1}$ to comove negatively. Thus, we have

$$\frac{\partial V^{(k)}(d_t, w_t)}{\partial w_t} < 0. \tag{A-60}$$

Turning to the second-order partial derivatives of $V^{(k)}(d_t, w_t)$:

1. We have

$$\frac{\partial^2 \mathcal{V}^{(k)}(d_t, w_t)}{\partial \left(d_t\right)^2} = 2\rho_d^2 \left(\mathrm{Var}_t \left[\frac{\partial s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)}{\partial d_{t+1}} \right] + \mathrm{Cov}_t \left[s^{(k-1)}\left(d_{t+1}, w_{t+1}\right), \frac{\partial^2 s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)}{\partial \left(d_{t+1}\right)^2} \right] \right).$$

   Since the variance term is positive, this expression will be positive unless the covariance term is negative and large. Furthermore, the covariance term depends on third derivatives of $s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)$ and these third order terms will vanish in continuous time as the time increment grows short. Thus, we assume it is safe to conclude that $\partial^2 \mathcal{V}^{(k)}(d_t, w_t)/\partial \left(d_t\right)^2 > 0$.

2. We have

$$\begin{aligned} \frac{\partial^2 \mathcal{V}^{(k)}(d_t, w_t)}{\partial d_t \partial w_t} &= 2\rho_d \rho_w \mathrm{Cov}_t \left[\frac{\partial s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)}{\partial w_{t+1}}, \frac{\partial s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)}{\partial d_{t+1}} \right] \\ &+ 2\rho_d \rho_w \mathrm{Cov}_t \left[s^{(k-1)}\left(d_{t+1}, w_{t+1}\right), \frac{\partial^2 s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)}{\partial d_{t+1} \partial w_{t+1}} \right]. \end{aligned}$$

   Under our inductive hypothesis the first covariance term will have the opposite sign as d_t. While the sign of the second covariance term is less obvious, it depends on the third derivatives of $s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)$ and these third order terms vanish in continuous time as the time increment grows short. Thus, we assume it is safe to conclude that $sgn\left(\partial^2 \mathcal{V}^{(k)}(d_t, w_t)/\partial d_t \partial w_t\right) = -sgn\left(d_t\right)$.

3. Finally, we have

$$\frac{\partial \mathcal{V}^{(k)}(d_t, w_t)}{\partial w_t} = 2\rho_w^2 \left(\mathrm{Var}_t \left[\frac{\partial s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)}{\partial w_{t+1}} \right] + \mathrm{Cov}_t \left[s^{(k-1)}\left(d_{t+1}, w_{t+1}\right), \frac{\partial^2 s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)}{\partial \left(w_{t+1}\right)^2} \right] \right).$$

   Since the variance term is positive, this expression will be positive unless the covariance term is negative and large. Furthermore, the covariance term depends on third derivatives of $s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)$ and these third order terms will vanish in continuous time as the time increment grows short. Thus, we assume it is safe to conclude that $\partial^2 \mathcal{V}^{(k)}(d_t, w_t)/\partial \left(w_t\right)^2 > 0$.

Characterizing $s^{(k)}\left(d_t, w_t\right)$: We have[48]

$$\begin{aligned} \frac{\partial s^{(k)}\left(d_t, w_t\right)}{\partial d_t} &= (1-\delta) \left[\begin{array}{c} (-\kappa_x)\left[sgn\left(d_t\right)\right]^2 \cdot \frac{\partial \psi(|d_t|, w_t)}{\partial |d_t|} \\ +(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2 \mathcal{V}^{(k)}(d_t, w_t) + (-\alpha)\left(\frac{\delta}{1-\delta}\right)^2 \frac{\partial \mathcal{V}^{(k)}(d_t, w_t)}{\partial d_t} d_t \end{array} \right] \\ &+ \delta \rho_d \mathrm{E}_t \left[\frac{\partial s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)}{\partial d_{t+1}} \right] < 0 \end{aligned}$$

and

$$\frac{\partial s^{(k)}\left(d_t, w_t\right)}{\partial w_t} = (1-\delta) \left[\begin{array}{c} (-\kappa_x) sgn\left(d_t\right) \cdot \frac{\partial \psi(|d_t|, w_t)}{\partial w_t} \\ +(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2 \frac{\partial \mathcal{V}^{(k)}(d_t, w_t)}{\partial w_t} \cdot d_t \end{array} \right] + \delta \rho_w \mathrm{E}_t \left[\frac{\partial s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)}{\partial w_{t+1}} \right].$$

Here the sign of the term in square brackets is the same as that of d_t. Moreover, we have $sgn\left(\partial s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)/\partial w_{t+1}\right) = sgn\left(d_{t+1}\right)$ and thus, we obtain $sgn\left(E_t\left[\partial s^{(k-1)}\left(d_{t+1}, w_{t+1}\right)/\partial w_{t+1}\right]\right) = sgn\left(d_t\right)$.

Turning to the second-order partial derivatives of $s^{(k)}(d_t, w_t)$:

[48] This expression omits the potential discrete jump down in $-\kappa_x sgn\left(d_t\right) \cdot \psi\left(w_t, |d_t|\right)$ that occurs as d_t crosses from negative to positive.

1. We have

$$\frac{\partial^2 s^{(k)}(d_t,w_t)}{\partial (d_t)^2} \overset{a.s.}{=} (1-\delta)(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2 \left(2\frac{\partial \mathcal{V}^{(k)}(w_t,d_t)}{\partial d_t} + \frac{\partial^2 \mathcal{V}^{(k)}(w_t,d_t)}{\partial (d_t)^2} d_t\right) + \delta\rho_d^2 \mathrm{E}_t\left[\frac{\partial^2 s^{(k-1)}(d_{t+1},w_{t+1})}{\partial (d_{t+1})^2}\right].$$

Since $sgn\left(\partial V^{(k)}(d_t,w_t)/\partial d_t\right) = sgn(d_t)$ and $\partial^2\mathcal{V}^{(k)}(d_t,w_t)/\partial (d_t)^2 > 0$, the sign of the term in the square brackets is the opposite as that of d_t. Further, $sgn(\partial^2 s^{(k-1)}(d_{t+1},w_{t+1})/\partial (d_{t+1})^2) = -sgn(d_{t+1})$. Therefore, we have

$$sgn\left(\frac{\partial^2 s^{(k)}(d_t,w_t)}{\partial (d_t)^2}\right) = -sgn(d_t).$$

2. We have

$$\frac{\partial^2 s^{(k)}(d_t,w_t)}{\partial d_t \partial w_t} \overset{a.s.}{=} (1-\delta)(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2 \left(\frac{\partial^2 \mathcal{V}^{(k)}(d_t,w_t)}{\partial w_t \partial d_t}\cdot d_t + \frac{\partial \mathcal{V}^{(k)}(d_t,w_t)}{\partial w_t}\right) + \delta\rho_d\rho_w \mathrm{E}_t\left[\frac{\partial^2 s^{(k-1)}(d_{t+1},w_{t+1})}{\partial w_{t+1}\partial d_{t+1}}\right]$$

Since $sgn\left(\partial^2\mathcal{V}^{(k)}(d_t,w_t)/\partial w_t\partial d_t\right) = -sgn(d_t)$ and $\partial V^{(k)}(d_t,w_t)/\partial w_t < 0$, the term in the square brackets is positive. Further, $\partial^2 s^{(k-1)}(d_{t+1},w_{t+1})/\partial w_{t+1}\partial d_{t+1} > 0$. Therefore, we have

$$\frac{\partial^2 s^{(k)}(d_t,w_t)}{\partial w_t \partial d_t} > 0.$$

3. Finally, we have

$$\frac{\partial^2 s^{(k)}(d_t,w_t)}{\partial (w_t)^2} \overset{a.s.}{=} (1-\delta)(-\alpha)\left(\frac{\delta}{1-\delta}\right)^2 \frac{\partial^2 \mathcal{V}^{(k)}(d_t,w_t)}{\partial (w_t)^2}\cdot d_t + \delta\rho_w^2 \mathrm{E}_t\left[\frac{\partial^2 s^{(k-1)}(d_{t+1},w_{t+1})}{\partial (w_{t+1})^2}\right] > 0.$$

Since $\partial^2\mathcal{V}^{(k)}(d_t,w_t)/\partial (w_t)^2 > 0$ and $sgn\left(\partial^2 s^{(k-1)}(d_{t+1},w_{t+1})/\partial (w_{t+1})^2\right) = -sgn(d_{t+1})$, we have

$$sgn\left(\frac{\partial^2 s^{(k)}(d_t,w_t)}{\partial (w_t)^2}\right) = -sgn(d_t).$$

Thus, we have shown that if properties (A-47a), (A-47b), (A-47c), (A-48a), and (A-48b) hold for $k-1$, then they also for k. It follows that these properties then also hold for $s_t = s(d_t,w_t) = \lim_{k\to\infty} s^{(k)}(d_t,w_t)$. ■

Proof of Proposition 10. The proof of Proposition 10 follows immediately from that of Proposition 11 since, under the assumptions of Proposition 10, we always have $sgn(d_t) = sgn(d_{t+k}) = 1$. ■

B.4 Term structure model

B.4.1 Model

Since it features a single long-term perpetual swap, our baseline model does not deliver specific implications for the full term structure of swap spreads. In order to study this term structure in a simple fashion, we introduce a series of n-period zero-coupon swaps alongside the perpetual swap. We then take the limit as net end-user demand for these zero-coupon swaps goes to zero. Using this approach, we can derive the spreads on these non-traded n-period zero-coupon swaps from intermediaries' first order conditions.

A position in the n-period swap spread trade receives the fixed rate on an n-period zero-coupon swap and hedges the associated interest rate risk by going short n-period zero-coupon Treasury bonds. As a result, the return on this n-period receive-fixed spread trade is

$$r_{t+1}^{s(n)} = ns_t^{(n)} - (n-1)\,s_{t+1}^{(n-1)} - m_t, \quad \text{(A-61)}$$

where $s_t^{(n)} \equiv y_t^{S(n)} - y_t^{T(n)}$—i.e., the n-period zero-coupon swap spread is defined as the between the n-period swap yield, $y_t^{S(n)}$, and the n-period Treasury yield, $y_t^{T(n)}$.[49]

General analysis. Intermediaries maximize

$$\max_{o_t, x_t, \{x_t^{(n)}\}_{n=1}^{N}} \mathrm{E}_t\left[w_{t,t+1}\right] - \frac{\alpha}{2}\mathrm{Var}_t\left[w_{t,t+1}\right], \tag{A-62}$$

subject to the budget constraint

$$w_{t,t+1} = w_t + x_t r_{t+1}^s + \sum\nolimits_{n=1}^{N} x_t^{(n)} r_{t+1}^{s(n)} + o_t r_{t+1}^o \tag{A-63}$$

and the leverage constraint

$$\kappa_x\left|x_t\right| + \kappa_x \sum\nolimits_{n=1}^{N} |x_t^{(n)}| + \kappa_o\left|o_t\right| \leq w_t. \tag{A-64}$$

In other words, we now allow intermediaries to take positions in the perpetual long-term swap as well as in n-period zero-coupon swaps for $n = 1, ..., N$.[50]

The first-order condition for the position in n-period swaps ($x_t^{(n)}$) is

$$\mathrm{E}_t[r_{t+1}^{s(n)}] = \alpha\mathrm{Cov}_t[r_{t+1}^{s(n)}, w_{t,t+1}] + \kappa_x sgn(x_t^{(n)}) \cdot \psi_t. \tag{A-65}$$

The first-order conditions for the position in the perpetual swap (x_t) and the outside investment opportunity (o_t) are

$$\mathrm{E}_t[r_{t+1}^s] = \alpha\mathrm{Cov}_t[r_{t+1}^s, w_{t,t+1}] + \kappa_x sgn\left(x_t\right) \cdot \psi_t \text{ and} \tag{A-66}$$

$$\bar{r}_o = \alpha\sigma_o^2 \cdot o_t + \kappa_o sgn\left(o_t\right) \cdot \psi_t. \tag{A-67}$$

Since $\bar{r}_o > 0$ and $\psi_t \geq 0$, we must have $o_t \geq 0$, and in turn we have

$$\psi_t = \psi(|x_t| + \sum\nolimits_{n=1}^{N} |x_t^{(n)}|, w_t) = \max\left\{0, \frac{\alpha\sigma_o^2}{\kappa_o^2}\left(\kappa_o\frac{\bar{r}_o}{\alpha\sigma_o^2} + \kappa_x\left(|x_t| + \sum\nolimits_{n=1}^{N} |x_t^{(n)}|\right) - w_t\right)\right\}.$$

To compute equilibrium swap spreads, we impose market clearing, setting $x_t = -d_t$ and $x_t^{(n)} = -d_t^{(n)}$ for all n. To close the model a parsimonious way, we assume the same demand process for the perpetual swap d_t introduced above in (7) and then take the limit as $d_t^{(n)}$ approaches zero from one side. Taking this one-sided limit as positions in the n-maturity swaps grow small is necessary to preserve the terms relating to balance sheet costs.[51]

Taking this limit, we obtain

$$\overbrace{ns_t^{(n)} - (n-1)\,E_t[s_{t+1}^{(n-1)}] - m_t}^{\mathrm{E}_t[r_{t+1}^{(n)}]} = -\alpha\mathrm{Cov}_t[r_{t+1}^{s(n)}, r_{t+1}^s]d_t - sgn(d_t^{(n)})\kappa_x\psi_t. \tag{A-68}$$

where $\mathcal{C}_t^{(n)} \equiv \mathrm{Cov}_t[r_{t+1}^{s(n)}, r_{t+1}^s] = (n-1)\,\frac{\delta}{1-\delta}\mathrm{Cov}_t[s_{t+1}^{(n-1)}, s_{t+1}]$ is the risk of n-period swaps captured by their return co-movement with the return on the long-maturity swap, and we have used the fact that under our assumptions the equilibrium Lagrange multiplier at time t is a function of only w_t and $|d_t|$, i.e., $\psi\left(|d_t|, w_t\right)$. This expression naturally generalizes equation (14). Iterating this equation forward, we obtain the following expression the n -period

[49]The return on an n-period zero coupon swap from t to $t+1$ is $r_{t+1}^{S(n)} \equiv ny_t^{S(n)} - (n-1)\,y_{t+1}^{S(n-1)} - i_t^S$ where $y_t^{S(n)}$ is the n-period zero-coupon swap yield at time t. Similarly, the excess return on an n-period zero-coupon Treasury is $r_{t+1}^{T(n)} \equiv ny_t^{T(n)} - (n-1)\,y_{t+1}^{T(n-1)} - i_t^T$ where $y_t^{T(n)}$ is the n-period zero-coupon Treasury yield at time t. Thus, the return on the n-period swap spread arbitrage trade takes the form given above.

[50]A simplifying assumption here is that all swap spread arbitrage positions have the same margin requirement κ_x irrespective of maturity n. Allowing these margin requirements to vary by maturity would introduce additional complexity without qualitatively changing our results.

[51]There are other ways we could close the model. Specifically, we could set $d_t \equiv 0$ and specify a demand process for the set of swap maturities $\{d_t^{(n)}\}_{n=1}^{N}$. This approach would allow us to consider separate shocks to the demand for swaps with different maturities. We have opted to close the model in a more parsimonious way here since we are primarily interested in how the general level of demand for long-term swaps impacts the term structure of swap spreads. However, below we will extend our model to allow for multiple demand factors.

swap spread

$$s_t^{(n)} = \overbrace{n^{-1}\sum_{k=0}^{n-1} \mathrm{E}_t[m_{t+k}]}^{\text{Expected short-rate differentials}} + \overbrace{n^{-1}\sum_{k=0}^{n-1} \mathrm{E}_t[(-\kappa_x)\, sgn(d_{t+k}^{(n-k)})\psi\left(|d_{t+k}|, w_{t+k}\right)]}^{\text{Expected compensation for using scarce capital}} + \underbrace{n^{-1}\sum_{k=0}^{n-1} \mathrm{E}_t[(-\alpha)\, \mathcal{C}_{t+k}^{(n-k)} d_{t+k}]}_{\text{Expected compensation for risk}}, \quad \text{(A-69)}$$

which generalizes equation (15).

Consider the spreads on 1-period swaps, $s_t^{(1)}$. Since the return on a 1-period swap spread trade is riskless, we have $\mathcal{C}_t^{(1)} = 0$ implying that

$$s_t^{(1)} = m_t + (-\kappa_x)\, sgn(d_t^{(1)})\psi\left(|d_t|, w_t\right). \quad \text{(A-70)}$$

Thus, the short-dated spread is the sum of the current short-rate differential (m_t) and a term that is proportional to the current shadow cost of intermediary capital ($\psi_t \geq 0$). If the net demand to receive fixed is always positive for all maturities n—i.e., $sgn(d_t^{(n)}) = sgn(d_t) = 1$ for all t and n, then $s_t^{(1)} = m_t - \kappa_x \psi_t$ and $s_t^{(n)} = n^{-1}\sum_{k=0}^{n-1} \mathrm{E}_t[s_{t+k}^{(1)}] + n^{-1}\sum_{k=0}^{n-1} \mathrm{E}_t[(-\alpha)\, \mathcal{C}_{t+k}^{(n-k)} d_{t+k}]$. In this case, the short-term swap spread $s_t^{(1)}$ plays a role that is analogous to that played by short-term interest rates in traditional term structure models.

Assume that $m_t \equiv 0$ for all t as is the case for a SOFR swap, so that any frictionless model would predict that $s_t^{(n)} = 0$ for all n and t by the LoOP. Naturally, this obtains in our model in the limit where $\kappa_x = 0$. However, when $\kappa_x > 0$, swap spreads will no longer be zero due to the potential for binding intermediary capital constraints. Indeed, the current short-term spread ($s_t^{(1)}$) reveals the current shadow value of intermediary capital (ψ_t) up to a constant of proportionality. If $m_t \equiv 0$, short-dated spreads will be zero at t only if the constraint is slack at t ($\psi_t = 0$).

Now consider the spreads on swaps with $n \geq 2$ periods and continue to assume that $m_t \equiv 0$ for all t. Crucially, $s_t^{(n)}$ can still be non-zero for $n \geq 2$ and will be impacted by supply and demand even when $\psi_t = 0$ today In order to have $s_t^{(n)} \neq 0$ for $n > 1$, we only need to have $\mathrm{E}_t[\psi_{t+k}] > 0$ for some $k \leq n-1$—i.e., the constraint must be expected to bind sometime over the life of the swap.

Furthermore, even if the constraint is slack at time t, the expected returns from t to $t+1$ for spread arbitrage on long-dated swaps will be non-zero since there can be news at $t+1$ about the severity of future constraints. Specifically, even if $\psi_t = 0$, we will have $\mathrm{E}_t[r_{t+1}^{s(n)}] = -\alpha \mathrm{Cov}_t[r_{t+1}^{s(n)}, r_{t+1}^{s}]d_t \neq 0$, so long as $\mathrm{Var}_t[\mathrm{E}_{t+1}[\psi_{t+k}]] > 0$ for some $k \leq n-1$. That is, even if the constraint is slack at t, intermediaries need to be compensated for the risk of swap spread movements at $t+1$ due to news about the tightness of future constraints. Intermediaries still need to be compensated for bearing demand-and-supply risk between t and $t+1$ even when the constraint is not currently binding.

An affine equilibrium. To obtain an affine model of the term structure of swap spreads, we conjecture that

$$s_t^{(n)} = A_0^{(n)} + A_m^{(n)} z_t^m + A_d^{(n)} z_t^d + A_w^{(n)} z_t^w, \quad \text{(A-71)}$$

and impose assumptions A1, A2, and A3 (if necessary). To obtain this solution, we take the upper limit as $d_t^{(n)}$ approaches zero from above for all n—i.e., we assume that intermediaries have a vanishingly small short position in the pay-fixed swap spread arbitrage for all maturities n. We obtain the following result:

Theorem 2 *The equilibrium n-period swap spread is given by*

$$s_t^{(n)} = A_0^{(n)} + A_m^{(n)} z_t^m + A_d^{(n)} z_t^d + A_w^{(n)} z_t^w, \quad \text{(A-72)}$$

with the exact functional forms of $A_0^{(n)}$, $A_m^{(m)}$, $A_d^{(n)}$, and $A_w^{(n)}$ provided in Appendix B.4.2.

When $\gamma = 0$, $A_0^{(n)}$ is negative and decreasing in maturity n, $A_m^{(n)}$ and $A_w^{(n)}$ are positive and decreasing in n, and $A_d^{(n)}$ is negative for all n. $A_d^{(n)}$ reflects both the (i) expected balance sheet costs and (ii) compensation for risk over the life of the swap. When the volatility of swap spreads is sufficiently low, (i) dominates and $A_d^{(n)}$ is an increasing function of maturity n. When the volatility of swap spreads is higher, (ii) dominates. In this case, $A_d^{(n)}$ is downward-sloping across maturities n when ρ_d is sufficiently high and is a U-shaped function of maturity n when ρ_d is lower.

As shown by (A-69), swap spreads equal the average expected short-rate differential plus the average expected returns to swap spread arbitrage over the lifetime of a swap. And, the expected returns to swap spread arbitrage reflect the expected compensation for using scarce capital and the expected compensation for risk.

For simplicity, we focus on the case where end-user demand is completely inelastic ($\gamma = 0$). A key implication of our model is that the swap spread curve is downward-sloping on average when end-user demand to receive fixed is positive ($d_t > 0$). This result arises because, at least when end-user demand is completely inelastic ($\gamma = 0$), $\mathcal{C}^{(n)} \equiv \text{Cov}_t[r_{t+1}^{s(n)}, r_{t+1}^s]$ is an increasing function of maturity n—i.e., the returns on longer-dated swaps covary more strongly with the returns on intermediaries' portfolios—implying that longer-dated swaps are riskier for intermediaries. The greater average magnitude of longer-dated spreads reflects the greater risk compensation that intermediaries require on these longer-dated swaps.

Turning to the level of spreads, when $\text{E}[m_t] = \overline{m} \geq 0$ is sufficiently small, such as for OIS swaps, we have $A_0^{(n)} < 0$ for all n. When $\overline{m}$ is sufficiently large, we have $A_0^{(n)} > 0$ for all n. In the intermediate case where $\overline{m}$ is moderately positive, we have $A_0^{(n)} > 0$ for small n and $A_0^{(n)} < 0$ for larger n. Thus, a higher interest rate differential pushes all swap spreads towards positive values, however, the balance sheet cost and the risk compensation terms push swap spreads towards negative levels. The relative magnitudes of these forces determine the overall sign of $A_0^{(n)}$, with longer maturity spreads more likely to be negative.

The $A_m^{(n)}$ term reflects variation in the expected future short rate differentials (m_t) over the life of the swap and, thus, is positive and locally downward-sloping across maturities n. Similarly, $A_w^{(n)}$ reflects variation in the impact of intermediary wealth on expected future balance sheet costs ($-\kappa_x \psi_t$) over the life of the swap and, thus, is positive and locally downward-sloping across maturities n.

Finally, $A_d^{(n)}$ is negative for all maturities n and reflects both the (i) expected balance sheet costs and (ii) compensation for risk over the life of the swap. When the volatility of swap spreads is sufficiently low, (i) dominates for all maturities, and $A_d^{(n)}$ is an increasing function of maturity n. When the volatility of swap spreads is higher, (ii) dominates. In this case, $A_d^{(n)}$ is downward-sloping across maturities n when ρ_d is sufficiently high and is a U-shaped function of maturity n when ρ_d is lower.

When $\gamma > 0$, end users submit price-sensitive demands and hence movements in z_t^m and z_t^w affect intermediaries' equilibrium exposure to swaps; this implies that $A_m^{(n)}$ and $A_w^{(n)}$ now also reflect compensation for risk. By continuity, all of the analytical results for the $\gamma = 0$ case continue to hold when end-user demand is somewhat inelastic, i.e., for $\gamma > 0$ sufficiently small. However, the local shape of the swap spread curve can become more complex when $\gamma > 0$ is larger, —i.e., when end-user demand is highly elastic, which creates strong feedback effects between current and expected future swap spreads and can theoretically lead to swap spread curves with oscillatory properties that seem less empirically realistic.

Expected returns also have a term structure, as follows:

Proposition 12 *The expected return on the n-period swap spread trade is given by*

$$E_t[r_{t+1}^{s(n)}] = B_0^{(n)} + B_m^{(n)} z_t^m + B_m^{(n)} z_t^d + B_w^{(n)} z_t^w, \tag{A-73}$$

with the exact functional forms of $B_0^{(n)}$, $B_m^{(m)}$, $B_d^{(n)}$, and $B_w^{(n)}$ given in Internet Appendix B.4.2.

When end-user demand is completely inelastic ($\gamma = 0$), (i) $B_0^{(n)}$ is negative and decreasing in maturity n, (ii) $B_m^{(n)} = 0$ for all n, (iii) $B_d^{(n)}$ is negative and decreasing in n, and (iv) $B_w^{(n)}$ is positive and constant across maturities.

When end-user demand is somewhat inelastic (γ is positive, but not too large), (i) $B_0^{(n)} < 0$, (ii) $B_m^{(n)} < 0$, (iii) $B_d^{(n)} < 0$, and (iv) $B_w^{(n)} > 0$. Since $\mathcal{C}^{(n)}$ is increasing in n when γ is not too large, all four coefficients are decreasing in maturity n in this case.

When $\gamma = 0$, interest rate differential shocks and supply shocks have the same effect on the expected swap arbitrage returns for all maturities n, as measured by $B_m^{(n)}$ and $B_w^{(n)}$, respectively. In fact, $B_m^{(n)} = 0$ for all maturities because, just like in the case of the perpetual long-term swap spread, changes in the short-rate differential m_t impact swap cash flows, but do not affect the equilibrium amount of swap spread risk that intermediaries must hold. In turn, the constancy of $B_w^{(n)}$ reflects the fact that as long as $\gamma = 0$, the $B_w^{(n)} z_t^w$ term in (A-73) purely reflects compensation for consuming scarce capital, and all swaps, irrespective of maturity, consume the same amount of capital per unit notional.[52] By contrast, the $B_d^{(n)} z_t^d$ term reflects both compensation for consuming scarce capital and compensation for risk. The fact that $B_d^{(n)}$ is decreasing in n reflects the fact that longer-term swaps are riskier for intermediaries—i.e., that $\mathcal{C}^{(n)}$ is increasing in n. The fact that longer-dated swaps are riskier also implies that the unconditional average of expected returns, $B_0^{(n)}$, must decline in maturity n.

When end-user demand is somewhat inelastic (γ is positive, but not too large), $B_0^{(n)}$, $B_m^{(n)}$, $B_d^{(n)}$, and $B_w^{(n)}$ are each decreasing in n. This decline reflects the facts that (i) $\text{E}_t[r_{t+1}^{s(n)}] = -\alpha \mathcal{C}^{(n)} d_t - \kappa_x \psi(w_t, d_t)$, (ii) $\mathcal{C}^{(n)}$ is increasing in n, and (iii) $d_t = \bar{d} + z_t^d + \gamma s_t$ is increasing in z_t^m, z_t^d, and z_t^w when $\gamma > 0$.

Using this term structure extension of our model, we can further study both the global as well as the local shape of the spread curve. Our measure of the global shape of the spread curve, $s_t - s_t^{(1)}$, is simply the difference in

[52] Formally, we have (i) $\text{E}_t[r_{t+1}^{Spread(n)}] = -\alpha \mathcal{C}^{(n)} d_t - \kappa_x \psi(w_t, d_t)$, (ii) $\mathcal{C}^{(n)}$ is increasing in n, and (iii) $d_t = \bar{d} + z_t^d$ is independent of z_t^w when $\gamma = 0$. Thus, $\partial \text{E}_t[r_{t+1}^{Spread(n)}]/\partial z_t^w = -\kappa_x \cdot \partial \psi_t / \partial z_t^w$ for all n.

spreads between the long-term generic swap—which we associate with a swap with the average duration of end-user demand—and a 1-period swap. Proposition 5 in the main text summarizes our results.[53]

We can also examine the local shape of the swap spread curve. We have:

Proposition 13 *Using (16) and (A-72), the* local *slope of the term structure of swap spreads is given by*

$$s_t^{(n)} - s_t^{(n-1)} = (A_0^{(n)} - A_0^{(n-1)}) + (A_m^{(n)} - A_m^{(n-1)})z_t^m + (A_d^{(n)} - A_d^{(n-1)})z_t^d + (A_w^{(n)} - A_w^{(n-1)})z_t^w. \quad \text{(A-74)}$$

When end-user demand is completely inelastic ($\gamma = 0$), we find that $A_0^{(n)} - A_0^{(n-1)}$, $A_m^{(n)} - A_m^{(n-1)}$, and $A_w^{(n)} - A_w^{(n-1)} < 0$ for all maturities n, while $A_d^{(n)} - A_d^{(n-1)}$ is ambiguous and depends on the amount of swap spread risk and the persistence of shocks to end-user demand (ρ_d). However, if swap spread risk is sufficiently high and demand shocks are sufficiently persistent, we have $A_d^{(n)} - A_d^{(n-1)} < 0$ for all n.

Our framework also implies that the slope of the swap curve can be used to predict return the returns to swap spread arbitrage. To this end, we calculate Fama-Bliss-style regression coefficients. Proposition 6 in the main part of the paper presents our results when using the global slope, and the following proposition present the result using the local slope.

Proposition 14 *If swap spread risk V is sufficiently large, a larger local slope of the swap spread curve predicts that longer-maturity swap spread trades will outperform short-maturity trades. Specifically, the slope coefficient from the forecasting regression*

$$r_{t+1}^{s(n)} - r_{t+1}^{s(n')} = a + b \cdot (s_t^{(n)} - s_t^{(n')}) + e_{t+1}. \quad \text{(A-75)}$$

satisfies $b > 0$ for any $n > n'$.

Finally, our framework allows to study swap spread volatility. In particular, we obtain the following result:

Proposition 15 *The conditional and unconditional volatilities of swap spreads are given by*

$$Var_t[s_{t+1}^{(n)}] = (A_m^{(n)})^2\sigma_m^2 + (A_d^{(n)})^2\sigma_d^2 + (A_w^{(n)})^2\sigma_w^2$$

and

$$Var[s_t^{(n)}] = (A_m^{(n)})^2\frac{\sigma_m^2}{1-\rho_m^2} + (A_d^{(n)})^2\frac{\sigma_d^2}{1-\rho_d^2} + (A_w^{(n)})^2\frac{\sigma_w^2}{1-\rho_w^2}.$$

When end-user demand is completely inelastic ($\gamma = 0$) or moderately inelastic (γ is not too large) and swap spread risk is meaningful (σ_m^2, σ_d^2, and σ_w^2 are large), (i) shocks to short-rate differentials and intermediary have a larger impact on the volatility of short-dated swaps (n small) whereas (ii) shocks to end-user demand have a larger impact on the volatility of long-dated swaps (n large).

Formally, this follows from the fact that $|A_m^{(n)}|$ and $|A_m^{(w)}|$ are decreasing in n, whereas $|A_d^{(n)}|$ is either decreasing or a hump-shaped function of n when there is meaningful swap spread risk. Intuitively, the changing drivers of swap spread volatility at different maturities arises from the fact that shocks to short-rate differentials and intermediary wealth primarily affect the term structure by changing the expected short-dated spread over the life of a swap. Since these shocks are mean reverting, this expectations-hypothesis-like channel plays an greater role in driving movements in short-dated swap spreads. By contrast, shocks to demand primarily affect the term structure by changing the expected compensation for risk over the life of a swap. This term-premium-like channel plays a greater role in driving movements in long-dated spreads.

B.4.2 Proofs and derivations

Proof of Theorem 2. To derive the solution to our affine term structure model, we start from (17) and (A-68) in the case when $d_t > 0$ and $d_t^{(n)} \searrow 0$, and thus $sgn(d_t^{(n)}) = sgn(d_t^{(n)}) = 1$. Combining these equations, we obtain

$$\overbrace{ns_t^{(n)} - (n-1)\,\mathrm{E}_t[s_{t+1}^{(n-1)}] - m_t}^{\mathrm{E}_t[r_{t+1}^{s(n)}]=B_0^{(n)}+B_m^{(n)}z_t^m+B_m^{(n)}z_t^d+B_w^{(n)}z_t^w} + \alpha\overbrace{\mathrm{Cov}_t[r_{t+1}^{s(n)}, r_{t+1}^s]}^{\mathcal{C}^{(n)}}d_t = \mathrm{E}_t[r_{t+1}^s] + \alpha V d_t = -\kappa_x\psi_t. \quad \text{(A-76)}$$

[53]Alternatively, we could define $Slope_t^* \equiv s_t^{(n)} - s_t^{(1)}$ as the difference between the n- and the 1-period swaps; while the exact coefficients and certain conditions change, our qualitative results presented in Proposition 5 remain the same. Further, we could also study the local shape of the term structure, i.e., the determinants of $Slope_t^{(n)} \equiv s_t^{(n)} - s_t^{(n-1)}$ for some n to obtain similar qualitative results.

Using our affine conjecture for the n-period swap spread (A-72) to express $\mathrm{E}_t[r_{t+1}^{s(n)}]$, combining it with (7) and (20), (A-76) implies

$$\begin{aligned}
&\left[nA_0^{(n)} - (n-1)\,A_0^{(n-1)} - \overline{m} + \alpha\left(\overline{d} + \gamma A_0\right)\mathcal{C}^{(n)}\right] + \left[nA_m^{(n)} - \rho_m\,(n-1)\,A_m^{(n-1)} - 1 + \alpha\gamma A_m \mathcal{C}^{(n)}\right]\cdot z_t^m \\
&\quad + \left[nA_d^{(n)} - \rho_d\,(n-1)\,A_d^{(n-1)} + \alpha\,(1+\gamma A_d)\,\mathcal{C}^{(n)}\right]\cdot z_t^d + \left[nA_w^{(n)} - \rho_w\,(n-1)\,A_w^{(n-1)} + \alpha\gamma A_w \mathcal{C}^{(n)}\right]\cdot z_t^w \\
&\quad = \left[A_0 - \overline{m} + \alpha V\left(\overline{d} + \gamma A_0\right)\right] + \left[\frac{1-\rho_m\delta}{1-\delta}A_m - 1 + \alpha\gamma V A_m\right]\cdot z_t^m \\
&\quad + \left[\frac{1-\rho_d\delta}{1-\delta}A_d + \alpha V\,(1+\gamma A_d)\right]\cdot z_t^d + \left[\frac{1-\rho_w\delta}{1-\delta}A_w + \alpha\gamma V A_w\right]\cdot z_t^w. \qquad \text{(A-77)}
\end{aligned}$$

Letting $D_j^{(n)} \equiv nA_j^{(n)}$ for $j \in \{0, m, d, w\}$ and noting that

$$\mathcal{C}^{(n)} \equiv \mathrm{Cov}_t[r_{t+1}^{s(n)}, r_{t+1}^s] = \frac{\delta}{1-\delta}\left(D_m^{(n-1)}A_m\sigma_m^2 + D_d^{(n-1)}A_d\sigma_d^2 + D_w^{(n-1)}A_w\sigma_w^2\right), \qquad \text{(A-78)}$$

we can write this system of recursive equations more compactly as

$$\begin{bmatrix} D_0^{(n)} \\ \mathbf{D}^{(n)} \end{bmatrix} = \mathbf{\Phi}_1 \begin{bmatrix} D_0^{(n)} \\ \mathbf{D}^{(n-1)} \end{bmatrix} + \mathbf{\Phi}_0, \qquad \text{(A-79)}$$

where $\mathbf{D}^{(n)} \equiv \left[D_m^{(n)}, D_d^{(n)}, D_w^{(n)}\right]'$,

$$\mathbf{\Phi}_1 = \begin{bmatrix} 1 & -\alpha\frac{\delta}{1-\delta}\sigma_m^2\left(\overline{d}+\gamma A_0\right)A_m & -\alpha\frac{\delta}{1-\delta}\sigma_d^2\left(\overline{d}+\gamma A_0\right)A_d & -\alpha\frac{\delta}{1-\delta}\sigma_w^2\left(\overline{d}+\gamma A_0\right)A_w \\ 0 & \rho_m - \alpha\gamma\frac{\delta}{1-\delta}\sigma_m^2 A_m^2 & -\alpha\gamma\frac{\delta}{1-\delta}\sigma_d^2 A_d A_m & -\alpha\gamma\frac{\delta}{1-\delta}\sigma_w^2 A_w A_m \\ 0 & -\alpha\frac{\delta}{1-\delta}\sigma_m^2 A_m\,(1+\gamma A_d) & \rho_d - \alpha\frac{\delta}{1-\delta}\sigma_d^2 A_d\,(1+\gamma A_d) & -\alpha\frac{\delta}{1-\delta}\sigma_w^2 A_w\,(1+\gamma A_d) \\ 0 & -\alpha\gamma\frac{\delta}{1-\delta}\sigma_m^2 A_m A_w & -\alpha\gamma\frac{\delta}{1-\delta}\sigma_d^2 A_d A_w & \rho_w - \alpha\gamma\frac{\delta}{1-\delta}\sigma_w^2 A_w^2 \end{bmatrix} \qquad \text{(A-80)}$$

and

$$\mathbf{\Phi}_0 = \begin{bmatrix} (1+\alpha\gamma V)\,A_0 + \alpha V\overline{d} \\ \left(\frac{1-\rho_m\delta}{1-\delta} + \alpha\gamma V\right)A_m \\ \left(\frac{1-\rho_d\delta}{1-\delta} + \alpha\gamma V\right)A_d + \alpha V \\ \left(\frac{1-\rho_w\delta}{1-\delta} + \alpha\gamma V\right)A_w \end{bmatrix}, \qquad \text{(A-81)}$$

and the initial conditions $D_0^{(0)} = D_m^{(0)} = D_d^{(0)} = D_w^{(0)} = 0$. The solution of this system of recursive equations is then

$$\begin{bmatrix} D_0^{(n)} \\ \mathbf{D}^{(n)} \end{bmatrix} = (\mathbf{I} - \mathbf{\Phi}_1)^{-1}(\mathbf{I} - \mathbf{\Phi}_1^n)\,\mathbf{\Phi}_0. \qquad \text{(A-82)}$$

For example, the 1-period swap rate at time t is given by

$$s_t^{(1)} = m_t - \kappa_x\psi_t = A_0^{(1)} + A_m^{(1)}z_t^m + A_d^{(1)}z_t^d + A_w^{(1)}z_t^w, \qquad \text{(A-83)}$$

where $[A_0^{(1)},\ A_m^{(1)},\ A_d^{(1)},\ A_w^{(1)}]' = [D_0^{(1)}, \mathbf{D}^{(1)}]' = \Phi_0$. We have $D_m^{(1)} = A_m^{(1)} > 0$, $D_d^{(1)} = A_d^{(1)} < 0$, and $D_w^{(1)} = A_w^{(1)} > 0$. Since $D_0^{(1)} = A_0^{(1)} = \overline{m} - \kappa_x \cdot \mathrm{E}[\psi_t]$, the sign of $A_0^{(1)}$ is ambiguous.

To understand how the $D_j^{(n)}$ coefficients behave as a function of n, note that (A-82) implies

$$\begin{bmatrix} D_0^{(n)} \\ \mathbf{D}^{(n)} \end{bmatrix} - \begin{bmatrix} D_0^{(n-1)} \\ \mathbf{D}^{(n-1)} \end{bmatrix} = \mathbf{\Phi}_1^{n-1}\mathbf{\Phi}_0, \qquad \text{(A-84)}$$

or, ignoring the constant term, $\mathbf{D}^{(n)} - \mathbf{D}^{(n-1)} = \hat{\mathbf{\Phi}}_1^{n-1}\mathbf{D}^{(1)}$ where $\hat{\mathbf{\Phi}}_1$ is the 3×3 obtained from $\mathbf{\Phi}_1$ by deleting the first row and the first column:

$$\hat{\mathbf{\Phi}}_1 = \begin{bmatrix} \rho_m - \alpha\gamma\frac{\delta}{1-\delta}\sigma_m^2 A_m^2 & -\alpha\gamma\frac{\delta}{1-\delta}\sigma_d^2 A_d A_m & -\alpha\gamma\frac{\delta}{1-\delta}\sigma_w^2 A_w A_m \\ -\alpha\frac{\delta}{1-\delta}\sigma_m^2 A_m\,(1+\gamma A_d) & \rho_d - \alpha\frac{\delta}{1-\delta}\sigma_d^2 A_d\,(1+\gamma A_d) & -\alpha\frac{\delta}{1-\delta}\sigma_w^2 A_w\,(1+\gamma A_d) \\ -\alpha\gamma\frac{\delta}{1-\delta}\sigma_m^2 A_m A_w & -\alpha\gamma\frac{\delta}{1-\delta}\sigma_d^2 A_d A_w & \rho_w - \alpha\gamma\frac{\delta}{1-\delta}\sigma_w^2 A_w^2 \end{bmatrix}. \qquad \text{(A-85)}$$

In the general case where $\gamma > 0$ is large, the behavior of $\mathbf{D}^{(n)}$ can be quite complex as we explain below. However, their behavior in the case when end-user demand is inelastic ($\gamma = 0$) is straightforward and intuitive.

Assume for now that $\gamma = 0$. Letting $\phi_d \equiv \rho_d - \alpha \frac{\delta}{1-\delta}\sigma_d^2 A_d > \rho_d > 0$, we have $\mathbf{D}^{(n)} - \mathbf{D}^{(n-1)} = \hat{\mathbf{\Phi}}_1^{n-1}\mathbf{D}^{(1)}$ with

$$\hat{\mathbf{\Phi}}_1^{n-1} = \left[\begin{array}{ccc} \rho_m^{n-1} & 0 & 0 \\ -\alpha \frac{\delta}{1-\delta}\sigma_m^2 A_m \cdot \frac{\phi_d^{n-1}-\rho_m^{n-1}}{\phi_d-\rho_m} & \phi_d^{n-1} & -\alpha \frac{\delta}{1-\delta}\sigma_w^2 A_w \cdot \frac{\phi_d^{n-1}-\rho_w^{n-1}}{\phi_d-\rho_w} \\ 0 & 0 & \rho_w^{n-1} \end{array} \right]. \tag{A-86}$$

It follows that

$$D_m^{(n)} - D_m^{(n-1)} = \rho_m^{n-1} \cdot D_m^{(1)} > 0, \tag{A-87}$$
$$D_w^{(n)} - D_w^{(n-1)} = \rho_w^{n-1} \cdot D_w^{(1)} > 0, \tag{A-88}$$

and

$$D_d^{(n)} - D_d^{(n-1)} = -\alpha \frac{\delta}{1-\delta}\sigma_m^2 A_m \cdot \frac{\phi_d^{n-1} - \rho_m^{n-1}}{\phi_d - \rho_m} \cdot D_m^{(1)} + \phi_d^{n-1} \cdot D_d^{(1)} - \alpha \frac{\delta}{1-\delta}\sigma_w^2 A_w \cdot \frac{\phi_d^{n-1} - \rho_w^{n-1}}{\phi_d - \rho_w} \cdot D_w^{(1)} < 0. \tag{A-89}$$

Thus, when $\gamma = 0$, $D_m^{(n)}$, $D_d^{(n)}$, and $D_w^{(n)}$ are monotonic functions of n with the same signs as A_m, A_d, and A_w. Using equation (A-78), it then follows that $\mathcal{C}^{(n)} > 0$ for $n \geq 2$ and is increasing in n. Naturally, $\mathcal{C}^{(1)} = 0$ since 1-period swap spread positions are riskless.

When $\gamma = 0$, we can provide closed-form expressions for $A_m^{(n)}$, $A_d^{(n)}$, and $A_w^{(n)}$. After some algebra, we obtain the following expressions for $A_m^{(n)}$ and $A_w^{(n)}$:

$$A_m^{(n)} = \frac{1}{n}\frac{1-\rho_m^n}{1-\rho_m} > 0, \tag{A-90}$$
$$A_w^{(n)} = \frac{1}{\kappa_x}\alpha\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 \cdot \frac{1}{n}\frac{1-\rho_w^n}{1-\rho_w} > 0. \tag{A-91}$$

Since $n^{-1}(1-\rho^n)/(1-\rho) = n^{-1}\sum_{k=0}^{n-1}\rho^k$ is decreasing in n for $\rho \in (0,1)$, it follows that $A_m^{(n)}$ and $A_w^{(n)}$are positive and decreasing functions of n.

Turning to $A_d^{(n)}$, we have

$$A_d^{(n)} = \theta_o \cdot \frac{1}{n}\frac{1-\phi_d^n}{1-\phi_d} + \theta_m \cdot S^{(n)}(\phi_d, \rho_m) + \theta_w \cdot S^{(n)}(\phi_d, \rho_w) < 0, \tag{A-92}$$

where

$$\theta_o = -\alpha\sigma_o^2\left(\frac{\kappa_x}{\kappa_o}\right)^2 < 0, \theta_m = -\alpha\sigma_m^2\frac{\delta}{1-\delta\rho_m} < 0, \theta_w = -\alpha\sigma_w^2\frac{\delta}{1-\rho_w\delta}\left(\frac{\theta_o}{\kappa_x}\right)^2 < 0, \tag{A-93}$$

and

$$S^{(n)}(a,b) \equiv \frac{1}{n}\sum_{k=1}^{n} a^{n-k}\frac{1-b^{k-1}}{1-b} = \begin{cases} \frac{1}{a-b}\frac{1}{n}\left(\frac{1-a^n}{1-a} - \frac{1-b^n}{1-b}\right) & \text{if } a \neq 1 \\ \frac{1}{a-b}\frac{1}{n}\left(\frac{1-a^n}{1-a} - \frac{1-b^n}{1-b}\right) & \text{if } a = 1 \end{cases} > 0 \tag{A-94}$$

for $a > 0$ and $b \in (0,1)$. It is easy to show that $S^{(n)}(a,b)$ is an increasing function of n when $a \geq 1$ and a hump-shaped function of n when $a < 1$.

In (A-92), the first term reflects a combination of (i) expected future compensation for using scarce capital and (ii) expected future compensation for risk due to shocks to z_t^d (recall that $\phi_d \equiv \rho_d - \alpha\frac{\delta}{1-\delta}\sigma_d^2 A_d > 0$). This term is only present when $\kappa_x > 0$, but exists even when $\sigma_m^2, \sigma_w^2 \to 0$. The second and third terms reflect expected future compensation for risk due to shocks to z_t^m and z_t^w (as amplified amplified by the risk of shocks to z_t^d) and are only present when $\sigma_m^2 > 0$ and $\sigma_w^2 > 0$, respectively.

Turning to the shape of these functions, $n^{-1}(1-\phi_d^n)/(1-\phi_d) = n^{-1}\sum_{k=0}^{n-1}\phi_d^k$ is a decreasing function of n when $\phi_d < 1$, increasing when $\phi_d > 1$, and constant when $\phi_d = 1$. Since $\rho_m, \rho_w \in (0,1)$, $S^{(n)}(\phi_d, \rho_i)$, $i = m, w$ are hump-shaped functions of n when $\phi_d < 1$ and increasing functions of n when $\phi_d \geq 1$.

Since $\phi_d \equiv \rho_d - \alpha\frac{\delta}{1-\delta}\sigma_d^2 A_d$, this means that all three terms tend to be increasing in magnitude in n when both ρ_d, σ_d^2, and $|A_d|$ are large and $\phi_d > 1$—i.e., when there are large and persistent shocks to end-user demand.

By contrast, when $\phi_d < 1$, the magnitude of the first term will decline with n and the second and third terms will be hump-shaped in n. When σ_m^2 and σ_w^2 are small, the first term dominates and the magnitude of $A_d^{(n)}$ declines with n. By contrast, when σ_m^2 and σ_w^2 are larger, the second and third terms dominate and the magnitude of $A_d^{(n)}$ is a hump-shaped function of n.

Letting $B_d^{(n)} \equiv \partial \mathrm{E}_t[r_{t+1}^{s(n)}]/\partial z_t^d = \theta_o - \alpha \cdot \mathcal{C}^{(n)}$, this hump shape stems from the fact that the loading of n-period swap spreads on demand shocks satisfies

$$A_d^{(n)} = n^{-1} \sum\nolimits_{k=1}^{n} \rho_d^{n-k} B_d^{(k)} = \underbrace{\theta_o \cdot \frac{1}{n} \frac{1-\rho_d^n}{1-\rho_d}}_{\substack{\text{Expected compensation} \\ \text{for using scarce capital}}} + \underbrace{\frac{1}{n} \sum\nolimits_{k=1}^{n} \rho_d^{n-k} \cdot \left(-\alpha \mathcal{C}^{(k)}\right)}_{\text{Expected compensation for risk}} < 0, \tag{A-95}$$

—i.e., the impact of a shift in demand on n-period swap spreads is an average of the expected balance sheet costs and expected risk premia (the $-\alpha \rho_d^{n-k} \mathcal{C}^{(k)}$ terms) over the life of the swap. The first term in (A-95) is negative and increasing in n. Since $\mathcal{C}^{(n)}$ is positive and increasing in n, the second term is negative and decreasing in n when ρ_d is sufficiently large and is a negative and U-shaped function of n when ρ_d is lower.

Overall, when swap spread risk is high and demand shocks are sufficiently persistent (ρ_d is large), the $A_d^{(n)}$ coefficients will rise in magnitude with n since the impact of demand on risk premia $|B_d^{(n)}|$ rises with n. However, when demand shocks are more transitory, the $A_d^{(n)}$ coefficients become U-shaped since these larger risk premia are only expected to persist for a short fraction over the life of the swap.

Finally, we consider the $A_0^{(n)}$ coefficients. Combining (A-84) and (A-90)-(A-92), we obtain

$$\begin{aligned} D_0^{(n)} - D_0^{(n-1)} = A_0^{(1)} + \eta_m \cdot & \overbrace{\left[\frac{1-\rho_m^{n-1}}{1-\rho_m} + \eta_o \cdot \frac{1}{\phi_d - \rho_m}\left(\frac{1-\phi_d^{n-1}}{1-\phi_d} - \frac{1-\rho_m^{n-1}}{1-\rho_m}\right)\right]}^{>0} \\ + \eta_d \cdot \frac{1-\phi_d^{n-1}}{1-\phi_d} + \eta_w \cdot & \overbrace{\left[\frac{1-\rho_w^{n-1}}{1-\rho_w} + \eta_o \cdot \frac{1}{\phi_d - \rho_w}\left(\frac{1-\phi_d^{n-1}}{1-\phi_d} - \frac{1-\rho_w^{n-1}}{1-\rho_w}\right)\right]}^{>0}, \end{aligned} \tag{A-96}$$

which, after some algebra, leads to

$$\begin{aligned} A_0^{(n)} = A_0^{(1)} + \left[\eta_d \cdot T^{(n)}(\phi_d)\right] + & \left[\eta_m \cdot \left(T^{(n)}(\rho_m) + \eta_o \cdot \frac{T^{(n)}(\phi_d) - T^{(n)}(\rho_m)}{\phi_d - \rho_m}\right)\right] \\ + & \left[\eta_w \left(T^{(n)}(\rho_w) + \eta_o \cdot \frac{T^{(n)}(\phi_d) - T^{(n)}(\rho_w)}{\phi_d - \rho_w}\right)\right], \end{aligned} \tag{A-97}$$

with

$$\begin{aligned} A_0^{(1)} &= \overline{m} - \kappa_x \cdot E[\psi_t] = \overline{m} - \kappa_x \frac{\alpha \sigma_o^2}{\kappa_o^2}\left[\kappa_o \frac{\overline{r}_o}{\alpha \sigma_o^2} + \kappa_x \overline{d} - \overline{w}\right], \\ \eta_o &= -\alpha \frac{\delta}{1-\delta} \sigma_d^2 A_d > 0, \eta_m = -\alpha \frac{\delta}{1-\delta} \sigma_m^2 \overline{d} A_m A_m^{(1)} < 0, \\ \eta_d &= -\alpha \frac{\delta}{1-\delta} \sigma_d^2 \overline{d} A_d A_d^{(1)} < 0, \eta_w = -\alpha \frac{\delta}{1-\delta} \sigma_w^2 \overline{d} A_w A_w^{(1)} < 0, \text{ and} \\ T^{(n)}(a) &\equiv \frac{1}{1-a}\left(1 - \frac{1}{n}\frac{1-a^n}{1-a}\right). \end{aligned} \tag{A-98}$$

$T^{(n)}(a)$ is positive and an increasing function of n for all $a > 0$, $a \neq 1$. $\left(T^{(n)}(a) - T^{(n)}(b)\right)/(a-b)$ is zero for $n = 1$ and otherwise positive and an increasing function of n for all a, $b > 0$ for which it is defined. Therefore, the three terms in square brackets in (A-97) are each negative and decreasing in n. It follows that $A_0^{(n)}$ is a decreasing function of n.

Turning to the sign of the $A_0^{(n)}$, note that the sign of $A_0^{(1)}$ is ambiguous and depends on $\overline{m} \geq 0$: $A_0^{(1)} < 0$ if $\overline{m}$ is sufficiently small. Thus, if $\overline{m} = 0$, as would be the case for SOFR swaps, $A_0^{(n)}$ is negative for all n. By contrast, as if $\overline{m}$ is a moderately small number, as could be the case for LIBOR swaps, $A_0^{(n)}$ will be positive for small n and negative for large n.

This completes our characterization of the solution when $\gamma = 0$. By continuity of the model's solution in γ, these same conclusions must also hold when $\gamma > 0$ is sufficiently near 0. However, these conclusions need not hold for γ large. Indeed, when the underlying shocks are transient and γ is very large, one can construct extreme parameterizations where $D_i^{(n)}$, $i = m, d, w$, and $\mathcal{C}^{(n)}$ are not monotonic and instead are oscillatory. Furthermore, the $A_i^{(n)}$ can also be oscillatory.

These oscillatory patterns are a quirk that arises from our modelling shortcut of working with zero-coupon bonds that are in infinitesimal supply. Specifically, these unintuitive oscillatory patterns for zero-coupon bonds obtain even though the coefficients for the perpetual swap have well-defined and intuitive limits when end-user demand becomes highly elastic—i.e., the A_i for $i = m, d, w$ converge to 0 as $\gamma \to \infty$. And, these oscillatory patterns for zero-coupon bonds would not arise if we closed the model in a more realistic—albeit mathematically less tractable—fashion where the zero-coupon bonds were available in non-zero supply. ■

Proof of Proposition 5. Combining (18b)-(19) with (A-81),

$$s_t - s_t^{(1)} = [A_0 - A_0^{(1)}] + [A_m - A_m^{(1)}]z_t^m + [A_d - A_d^{(1)}]z_t^d + [A_w - A_w^{(1)}]z_t^w,$$

where

$$\begin{bmatrix} A_0 - A_0^{(1)} \\ A_m - A_m^{(1)} \\ A_d - A_d^{(1)} \\ A_w - A_w^{(1)} \end{bmatrix} = \begin{bmatrix} -\alpha\left(\bar{d} + \gamma A_0\right) V < 0 \\ -\left[\frac{\delta}{1-\delta}\left(1-\rho_m\right) + \alpha\gamma V\right] A_m < 0 \\ -\left[\frac{\delta}{1-\delta}\left(1-\rho_d\right) + \alpha\gamma V\right] A_d - \alpha V \\ -\left[\frac{\delta}{1-\delta}\left(1-\rho_w\right) + \alpha\gamma V\right] A_w < 0 \end{bmatrix}.$$

From (A-81) and (A-83), it is easy to see that $\mathrm{E}[s_t - s_t^{(1)}] = A_0 - A_0^{(1)} = -\alpha V\, \mathrm{E}\left[d_t\right] < 0$, implying that the term structure of swap spreads—summarized by the difference between the generic long-term spread s_t and the 1-period spread—is downward-sloping on average. The average slope is a function of intermediaries' risk aversion α, swap spread risk V, and the average demand from end-users to receive the fixed swap rate, $\mathrm{E}\left[d_t\right] = \bar{d} + \gamma A_0 > 0$.

Using the expression for A_d, simple algebra shows that $A_d - A_d^{(1)}$ is negative if and only if (30) holds—i.e., if

$$\frac{\delta}{1-\delta}\left(1-\rho_d\right)\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 < V,$$

which is more likely when $\left(\kappa_x/\kappa_o\right)^2 \sigma_o^2$ is small and when ρ_d, σ_m^2, σ_d^2, and σ_w^2 are large. When this condition is met, $\mathrm{E}[s_t - s_t^{(1)}] < 0$ and $(s_t - s_t^{(1)})$ is decreasing in z_t^m, z_t^d, and z_t^w. ■

Proof of Proposition 6. First, note that $\mathrm{E}_t[r_{t+1}^s] = -\kappa_x \cdot \psi_t - \alpha V \cdot d_t$ and $\mathrm{E}_t[r_{t+1}^{s(1)}] = -\kappa_x \cdot \psi_t$, implying that

$$\begin{aligned} \mathrm{E}_t[r_{t+1}^s - r_{t+1}^{s(1)}] &= -\alpha V \cdot d_t = -\alpha V\left(\bar{d} + z_t^d + \gamma s_t\right) \\ &= -\alpha\left(\bar{d} + \gamma A_0\right) V - \alpha\gamma A_m V \cdot z_t^m - \alpha\left(1 + \gamma A_d\right) V \cdot z_t^d - \alpha\gamma A_w V \cdot z_t^w. \end{aligned}$$

Combining this with Proposition 5 assuming that condition (30) holds, it follows that both $\mathrm{E}_t[r_{t+1}^s - r_{t+1}^{s(1)}]$ and $s_t - s_t^{(1)}$ are decreasing in z_t^m, z_t^d, and z_t^w. And since since these three factors are orthogonal, this then implies that we must have $b > 0$ in the time-series regression (31). ■

Proof of Proposition 12. Note that the relationship between the $B_i^{(n)}$ and the $D_i^{(n)}$, $i \in \{0, m, d, w\}$, functions are given by

$$\begin{aligned} B_0^{(n)} &= D_0^{(n)} - D_0^{(n-1)} - \bar{m} \\ &= B_0 + \alpha\left(\bar{d} + \gamma A_0\right) \cdot (V - \mathcal{C}^{(n)}) = B_0^{(1)} - \alpha\left(\bar{d} + \gamma A_0\right) \cdot \mathcal{C}^{(n)} \\ B_m^{(n)} &= D_m^{(n)} - \rho_m D_m^{(n-1)} - 1 \\ &= B_m + \alpha\gamma A_m \cdot (V - \mathcal{C}^{(n)}) = B_m^{(1)} - \alpha\gamma A_m \cdot \mathcal{C}^{(n)} \\ B_d^{(n)} &= D_d^{(n)} - \rho_d D_d^{(n-1)} \\ &= B_d + \alpha\left(1 + \gamma A_d\right) \cdot (V - \mathcal{C}^{(n)}) = B_d^{(1)} - \alpha\left(1 + \gamma A_d\right) \cdot \mathcal{C}^{(n)} \\ B_w^{(n)} &= D_w^{(n)} - \rho_w D_w^{(n-1)} \\ &= B_w + \alpha\gamma A_w \cdot (V - \mathcal{C}^{(n)}) = B_w^{(1)} - \alpha\gamma A_w \cdot \mathcal{C}^{(n)} \end{aligned}$$

In the $\gamma = 0$ case, this reduces to $B_m^{(n)} = B_m = 0$ and $B_w^{(n)} = B_w > 0$. Furthermore, since $B_0^{(1)}, B_d^{(1)} < 0$ and $\mathcal{C}^{(n)}$ is increasing in n, both $B_0^{(n)}$ and $B_d^{(n)}$ are negative and decreasing in n.

When γ is positive but not too large, we can show that $B_0^{(n)} < 0$, $B_m^{(n)} < 0$, $B_d^{(n)} < 0$, and $B_w^{(n)} > 0$ and that each of these functions is decreasing in n. This follows from signing these functions for $n = 1$ and then using the fact that, by continuity, $\mathcal{C}^{(n)}$ is increasing in n when γ is not too large. Furthermore, when γ is sufficiently small, $B_w^{(n)}$ will remain positive even as n increases. ■

Proof of Proposition 13. For the local slope of the swap spread curve, we are interested in

$$s_t^{(n)} - s_t^{(n-1)} = \left[A_0^{(n)} - A_0^{(n-1)}\right] + \left[A_m^{(n)} - A_m^{(n-1)}\right] z_t^m + \left[A_d^{(n)} - A_d^{(n-1)}\right] z_t^d + \left[A_w^{(n)} - A_w^{(n-1)}\right] z_t^w.$$

Focusing on the $\gamma = 0$ case, using (A-90)-(A-92), we obtain

$$\begin{aligned} A_m^{(n)} - A_m^{(n-1)} &= \frac{1}{n}\frac{1-\rho_m^n}{1-\rho_m} - \frac{1}{n-1}\frac{1-\rho_m^{n-1}}{1-\rho_m} < 0, \\ A_w^{(n)} - A_w^{(n-1)} &= \frac{1}{\kappa_x}\alpha\left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 \cdot \left[\frac{1}{n}\frac{1-\rho_w^n}{1-\rho_w} - \frac{1}{n-1}\frac{1-\rho_w^{n-1}}{1-\rho_w}\right] < 0, \end{aligned} \tag{A-99}$$

and

$$\begin{aligned} A_d^{(n)} - A_d^{(n-1)} = \theta_o \cdot \left[\frac{1}{n}\frac{1-\phi_d^n}{1-\phi_d} - \frac{1}{n-1}\frac{1-\phi_d^{n-1}}{1-\phi_d}\right] + \theta_m \cdot \left[S^{(n)}(\phi_d,\rho_m) - S^{(n-1)}(\phi_d,\rho_m)\right] \\ + \theta_w \cdot \left[S^{(n)}(\phi_d,\rho_w) - S^{(n-1)}(\phi_d,\rho_w)\right], \end{aligned} \tag{A-100}$$

where $\theta_o, \theta_m, \theta_w < 0$. The first term in square brackets is positive when $\phi_d > 1$ and is negative when $\phi_d < 1$. When $\phi_d > 1$, the second and third terms in brackets are positive. However, when $\phi_d < 1$, second and third terms are positive for small n and negative for large n.

Thus, if $\phi_d \equiv \rho_d - \alpha\frac{\delta}{1-\delta}\sigma_d^2 A_d > 1$, we have $A_d^{(n)} - A_d^{(n-1)} < 0$ for all n. And, we have $\phi_d > 1$ if

$$V > \frac{1}{\alpha^2\sigma_d^2}\frac{1-\delta}{\delta}(1-\rho_d) - \left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2,$$

—e.g., if V is sufficiently large and ρ_d is sufficiently large. By contrast, if $\phi_d < 1$, the sign of $A_d^{(n)} - A_d^{(n-1)}$ depends on the magnitudes of σ_m^2, σ_d^2, and σ_w^2 and can depend on n. For instance, if σ_m^2 and σ_w^2 are small, the first term dominates and we have $A_d^{(n)} - A_d^{(n-1)} > 0$ for all n when $\phi_d < 1$. By contrast, if σ_m^2 and σ_w^2 are larger, the second and third terms dominate and we will have $A_d^{(n)} - A_d^{(n-1)} < 0$ for small n and $A_d^{(n)} - A_d^{(n-1)} > 0$ for large n. ■

Proof of Proposition 14. Note that $\mathrm{E}_t[r_{t+1}^{s(n)}] = -\kappa_x \cdot \psi_t - \alpha \cdot \mathcal{C}^{(n)} d_t$. For $n > n'$, this implies

$$\begin{aligned} \mathrm{E}_t[r_{t+1}^{s(n)} - r_{t+1}^{s(n')}] &= -\alpha\left(\mathcal{C}^{(n)} - \mathcal{C}^{(n')}\right) d_t \\ &= -\alpha\left(\mathcal{C}^{(n)} - \mathcal{C}^{(n')}\right)\left[(\bar{d} + \gamma A_0) + \overbrace{\gamma A_m}^{>0} \cdot z_t^m + \overbrace{(1+\gamma A_d)}^{>0} \cdot z_t^d + \overbrace{\gamma A_w}^{>0} \cdot z_t^w\right]. \end{aligned}$$

Here, when $\gamma = 0$, $\mathcal{C}^{(n)}$ is positive and increasing in n and thus $\mathcal{C}^{(n)} - \mathcal{C}^{(n')} > 0$, which also extends to the case where γ is small and positive by continuity. Therefore, $\mathrm{E}_t[r_{t+1}^{s(n)} - r_{t+1}^{s(n')}]$ is decreasing in z_t^m, z_t^d, and z_t^w. At the same time, Proposition 13 states that $s_t^{(n)} - s_t^{(n')}$ is always decreasing in z_t^m and z_t^w. However, the sign of its loading on z_t^d, $A_d^{(n)} - A_d^{(n')}$, is ambiguous. However, assuming that swap spread volatility is sufficiently high relative to the persistence of demand shocks, we have $A_d^{(n)} - A_d^{(n')} < 0$. In this case, we must have $b > 0$ in the time-series regression (A-75). ■

B.5 Model with multiple demand factors

B.5.1 Model

We now develop a simple model with two generic perpetual swaps—one with shorter duration and one with longer duration—to understand how our analysis changes if there are separate shocks to end-user demand for short-term and long-term swaps.

Following our original model, we write the return on short-dated swap spread trade

$$rx_{t+1}^{s,S} = \left(s_t^S - m_t\right) - \frac{\delta_S}{1-\delta_S}\left(s_{t+1}^S - s_t^S\right), \tag{A-101}$$

and the return on long-dated swap spread trade as

$$rx_{t+1}^{s,L} = \left(s_t^L - m_t\right) - \frac{\delta_L}{1-\delta_L}\left(s_{t+1}^L - s_t^L\right). \tag{A-102}$$

We assume that $\delta_L > \delta_S$, so the long-dated swap spread trade indeed has a greater duration.

At each date, intermediaries have mean-variance utility over 1-period ahead wealth. Let $\mathbf{x}_t = [x_t^S, x_t^L]'$ denote their position in the two receive-fixed swap spread trades, o_t their position in the outside investment opportunity, and $\mathbf{rx}_{t+1} = [rx_{t+1}^{s,S}, rx_{t+1}^{s,L}]'$ the vector of swap spread trade excess returns. Thus, at time t, intermediaries solve

$$\max_{\mathbf{x}_t} \left\{ \mathbf{x}_t' \mathrm{E}_t \left[\mathbf{rx}_{t+1} \right] - \frac{\alpha}{2} \left(\mathbf{x}_t \right)' \mathrm{Var}_t \left[\mathbf{rx}_{t+1} \right] \mathbf{x}_t + o_t \overline{r}_o - \frac{\alpha}{2} o_t^2 \sigma_o^2 \right\},$$

subject to the leverage constraint

$$\kappa_x \left| \mathbf{x}_t' \right| \mathbf{1}_2 + \kappa_o \left| o_t \right| \leq w_t.$$

The Lagrangian then becomes

$$\mathcal{L} = \mathbf{x}_t' \mathrm{E}_t \left[\mathbf{rx}_{t+1} \right] - \frac{\alpha}{2} \left(\mathbf{x}_t \right)' \mathrm{Var}_t \left[\mathbf{rx}_{t+1} \right] \mathbf{x}_t + o_t \overline{r}_o - \frac{\alpha}{2} o_t^2 \sigma_o^2 + \psi_t \left\{ w_t - \kappa_x \left| \mathbf{x}_t \right|' \mathbf{1}_2 - \kappa_o \left| o_t \right| \right\},$$

with first-order conditions

$$\mathrm{E}_t \left[\mathbf{rx}_{t+1} \right] = \alpha \mathrm{Var}_t \left[\mathbf{rx}_{t+1} \right] \mathbf{x}_t + \psi_t \kappa_x sgn(\mathbf{x}_t) \;\text{ and }\; \overline{r}_o = \alpha o_t \sigma_o^2 + \psi_t \kappa_o sgn(o_t). \tag{A-103}$$

As before, assuming $\overline{r}_o, \alpha \sigma_o^2 > 0$, we must have $o_t > 0$. Further, assuming that the constraint always binds—that i.e., $\psi_t > 0$ and $w_t = \kappa_x \left| \mathbf{x}_t' \right| \mathbf{1}_2 + \kappa_o \left| o_t \right|$, we obtain

$$\psi_t = \frac{\alpha \sigma_o^2}{\kappa_o} \left(\frac{\overline{r}_o}{\alpha \sigma_o^2} - o_t \right) = \frac{\alpha \sigma_o^2}{\kappa_o^2} \left(\kappa_o \frac{\overline{r}_o}{\alpha \sigma_o^2} + \kappa_x \left| \mathbf{x}_t \right|' \mathbf{1}_2 - w_t \right).$$

Together with (A-103), we then have

$$\mathrm{E}_t \left[\mathbf{rx}_{t+1} \right] = \alpha \mathrm{Var}_t \left[\mathbf{rx}_{t+1} \right] \mathbf{x}_t + sgn(\mathbf{x}_t) \kappa_x \frac{\alpha \sigma_o^2}{\kappa_o^2} \left(\kappa_0 \frac{\overline{r}_o}{\alpha \sigma_o^2} - w_t \right) + \alpha \left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 [sgn(\mathbf{x}_t) sgn(\mathbf{x}_t')] \mathbf{x}_t. \tag{A-104}$$

We assume that net end-user demand to receive the fixed rate for short- and long-dated swaps is

$$\overbrace{\begin{bmatrix} d_t^S \\ d_t^L \end{bmatrix}}^{\mathbf{d}_t} = \begin{bmatrix} \overline{d}^S + z_t^{d,S} + \gamma_S s_t^S \\ \overline{d}^L + z_t^{d,L} + \gamma_L s_t^L \end{bmatrix} = \overline{\mathbf{d}} + \mathbf{E}_d \mathbf{z}_t + \mathbf{\Gamma} \mathbf{s}_t, \tag{A-105}$$

where $\mathbf{z}_t = \left[z_t^m, z_t^{d,S}, z_t^{d,L}, z_t^w \right]'$ and

$$\mathbf{E}_d = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \text{ and } \mathbf{\Gamma} = \begin{bmatrix} \gamma_S & 0 \\ 0 & \gamma_L \end{bmatrix}.$$

Finally, we assume an autoregressive process for the four state variables governing the short rate differential (z_t^m), short- and long-maturity end-user demand ($z_t^{d,S}$ and $z_t^{d,L}$), and intermediary wealth (z_t^w). Specifically, we assume the vector of state variables $\mathbf{z}_t = [z_t^m, z_t^{d,S}, z_t^{d,L}, z_t^w]'$ follows

$$\mathbf{z}_{t+1} = \boldsymbol{\varrho} \mathbf{z}_t + \boldsymbol{\varepsilon}_{t+1}, \tag{A-106}$$

where $\boldsymbol{\varrho} = diag\left(\rho_m, \rho_{d,S}, \rho_{d,L}, \rho_w \right)$ is a diagonal matrix of AR(1) coefficients $\rho_m, \rho_{d,S}, \rho_{d,L}, \rho_w, \in [0, 1)$ and $\boldsymbol{\varepsilon}_{t+1} = [\varepsilon_{t+1}^m, \varepsilon_{t+1}^{d,S}, \varepsilon_{t+1}^{d,L}, \varepsilon_{t+1}^w]'$ is the vector of structural shocks. For simplicity, we assume the four structural shocks are orthogonal to each other: $\mathrm{Var}_t[\boldsymbol{\varepsilon}_{t+1}] \equiv \mathbf{\Sigma} = diag\left(\sigma_m^2, \sigma_{d,S}^2, \sigma_{d,L}^2, \sigma_w^2 \right)$ where $\sigma_i^2 \equiv \mathrm{Var}\left[\varepsilon_{t+1}^i \right]$ for $i = m, (d, S), (d, L), w$.

We conjecture an affine equilibrium in which equilibrium swap spreads are in the form

$$\overbrace{\begin{bmatrix} s_t^S \\ s_t^L \end{bmatrix}}^{\mathbf{s}_t} = \overbrace{\begin{bmatrix} A_0^S \\ A_0^L \end{bmatrix}}^{\mathbf{a}_0} + \overbrace{\begin{bmatrix} A_m^S & A_{d,S}^S & A_{d,L}^S & A_w^S \\ A_m^L & A_{d,s}^L & A_{d,L}^L & A_w^L \end{bmatrix}}^{\mathbf{A}_1 \equiv [\mathbf{a}_m, \mathbf{a}_{d,S}, \mathbf{a}_{d,L}, \mathbf{a}_w]} \overbrace{\begin{bmatrix} z_t^m \\ z_t^{d,S} \\ z_t^{d,L} \\ z_t^w \end{bmatrix}}^{\mathbf{z}_t},$$

which, together with (A-105), implies

$$\mathbf{d}_t = \overline{\mathbf{d}} + \mathbf{E}_d \mathbf{z}_t + \mathbf{\Gamma} \mathbf{s}_t = \left(\overline{\mathbf{d}} + \mathbf{\Gamma} \mathbf{a}_0 \right) + \left(\mathbf{E}_d + \mathbf{\Gamma} \mathbf{A}_1 \right) \mathbf{z}_t.$$

Finally, let

$$\mathbf{D}_{\boldsymbol{\delta}} = \begin{bmatrix} \frac{\delta_s}{1-\delta_S} & 0 \\ 0 & \frac{\delta_L}{1-\delta_L} \end{bmatrix}.$$

Combining the conjectured forms with definitions (A-101)-(A-102) and the AR(1) structure for the state variables, we obtain that excess returns are

$$\mathbf{rx}_{t+1} = \mathbf{s}_t - m_t \mathbf{1}_2 - \mathbf{D}_{\boldsymbol{\delta}} \left(\mathbf{s}_{t+1} - \mathbf{s}_t\right) = \left[\mathbf{a}_0 - \overline{m}\mathbf{1}\right] + \left[\mathbf{A}_1 + \mathbf{D}_{\boldsymbol{\delta}} \mathbf{A}_1 \left(\mathbf{I}_4 - \boldsymbol{\varrho}\right) - \mathbf{1}_2 \mathbf{e}_m'\right] \mathbf{z}_t - \mathbf{D}_{\boldsymbol{\delta}} \mathbf{A}_1 \boldsymbol{\varepsilon}_{t+1},$$

where $\mathbf{I}_4 \equiv diag\,(1,1,1,1)$ is the 4×4 identity matrix, $\mathbf{1}_2' = [1,1]$, and $\mathbf{e}_m' = [1,0,0,0]$. Thus, the first two moments of excess returns are

$$\begin{aligned} \mathrm{E}_t\left[\mathbf{rx}_{t+1}\right] &= \left[\mathbf{a}_0 - \overline{m}\mathbf{1}_2\right] + \left[\mathbf{A}_1 + \mathbf{D}_{\boldsymbol{\delta}} \mathbf{A}_1 \left(\mathbf{I}_4 - \boldsymbol{\varrho}\right) - \mathbf{1}_2 \mathbf{e}_m'\right] \mathbf{z}_t \text{ and} \\ \mathrm{Var}_t\left[\mathbf{rx}_{t+1}\right] &= \mathbf{D}_{\boldsymbol{\delta}} \mathbf{A}_1 \boldsymbol{\Sigma} \mathbf{A}_1' \mathbf{D}_{\boldsymbol{\delta}}' \equiv \mathbf{V} \equiv \begin{bmatrix} V_S & C_{S,L} \\ C_{S,L} & V_L \end{bmatrix}, \end{aligned} \tag{A-107}$$

where

$$\begin{aligned} V_S &= \left(\frac{\delta_s}{1-\delta_S}\right)^2 \left[\left(A_m^S\right)^2 \sigma_m^2 + \left(A_{d,S}^S\right)^2 \sigma_{d,S}^2 + \left(A_{d,L}^S\right)^2 \sigma_{d,L}^2 + \left(A_w^S\right)^2 \sigma_w^2\right], \\ C_{S,L} &= \frac{\delta_s}{1-\delta_S} \frac{\delta_L}{1-\delta_L} \left[A_m^S A_m^L \sigma_m^2 + A_{d,S}^S A_{d,S}^L \sigma_{d,S}^2 + A_{d,L}^S A_{d,L}^L \sigma_{d,L}^2 + A_w^S A_w^L \sigma_w^2\right], \text{ and} \\ V_L &= \left(\frac{\delta_L}{1-\delta_L}\right)^2 \left[\left(A_m^L\right)^2 \sigma_m^2 + \left(A_{d,S}^L\right)^2 \sigma_{d,S}^2 + \left(A_{d,L}^L\right)^2 \sigma_{d,L}^2 + \left(A_w^L\right)^2 \sigma_w^2\right]. \end{aligned}$$

Imposing market clearing, $\mathbf{x}_t = -\mathbf{d}_t$, (A-104) then becomes

$$\mathrm{E}_t\left[\mathbf{rx}_{t+1}\right] = -\alpha \mathbf{V} \mathbf{d}_t - sgn(\mathbf{d}_t) \kappa_x \frac{\alpha \sigma_o^2}{\kappa_o^2} \left(\kappa_0 \frac{\overline{r}_o}{\alpha \sigma_o^2} - w_t\right) - \alpha \sigma_o^2 \left(\frac{\kappa_x}{\kappa_o}\right)^2 \left[sgn(\mathbf{d}_t) sgn(\mathbf{d}_t')\right] \mathbf{d}_t,$$

which, when imposing $\mathbf{d}_t > 0$ and thus $sgn(\mathbf{d}_t) = \mathbf{1}_2$, simplifies to

$$\mathrm{E}_t\left[\mathbf{rx}_{t+1}\right] = -\alpha \mathbf{V} \mathbf{d}_t - \kappa_x \frac{\alpha \sigma_o^2}{\kappa_o^2} \left(\kappa_0 \frac{\overline{r}_o}{\alpha \sigma_o^2} - w_t\right) \mathbf{1}_2 - \alpha \sigma_o^2 \frac{\kappa_x^2}{\kappa_o^2} \mathbf{1}_2 \mathbf{1}_2' \mathbf{d}_t. \tag{A-108}$$

We can further rewrite the right-hand side as

$$\begin{aligned} \mathrm{E}_t\left[\mathbf{rx}_{t+1}\right] &= -\alpha \left[\mathbf{V} + \left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 \mathbf{1}_2 \mathbf{1}_2'\right] \left(\left(\overline{\mathbf{d}} + \boldsymbol{\Gamma} \mathbf{a}_0\right) + \left(\mathbf{E}_d + \boldsymbol{\Gamma} \mathbf{A}_1\right) \mathbf{z}_t\right) \\ &\quad - \kappa_x \frac{\alpha \sigma_o^2}{\kappa_o^2} \left(\kappa_0 \frac{\overline{r}_o}{\alpha \sigma_o^2} - \left(\overline{w} + \mathbf{e}_w' \mathbf{z}_t\right)\right) \mathbf{1}_2 \\ &= \left[-\alpha \left[\mathbf{V} + \left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 \mathbf{1}_2 \mathbf{1}_2'\right] \left(\overline{\mathbf{d}} + \boldsymbol{\Gamma} \mathbf{a}_0\right) - \kappa_x \frac{\alpha \sigma_o^2}{\kappa_o^2} \left(\kappa_0 \frac{\overline{r}_o}{\alpha \sigma_o^2} - \overline{w}\right) \mathbf{1}_2\right] \\ &\quad + \left[-\alpha \left[\mathbf{V} + \left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 \mathbf{1}_2 \mathbf{1}_2'\right] \left(\mathbf{E}_d + \boldsymbol{\Gamma} \mathbf{A}_1\right) + \kappa_x \frac{\alpha \sigma_o^2}{\kappa_o^2} \mathbf{1}_2 \mathbf{e}_w'\right] \mathbf{z}_t, \end{aligned} \tag{A-109}$$

where $\mathbf{e}_w' = [0,0,0,1]$. From here, matching the constant and the coefficients of $\mathbf{z}_t$ in (A-107) and (A-109), we must have

$$\begin{aligned} &\left[\mathbf{I}_4 + \alpha \left[\mathbf{V} + \left(\frac{\kappa_x}{\kappa_o}\right)^2 \sigma_o^2 \mathbf{1}_2 \mathbf{1}_2'\right] \boldsymbol{\Gamma}\right] \mathbf{a}_0 \\ &= \overline{m}\mathbf{1}_2 - \kappa_x \frac{\alpha \sigma_o^2}{\kappa_o^2} \left[\kappa_x \mathbf{1}_2 \mathbf{1}_2' \overline{\mathbf{d}} + \left(\kappa_0 \frac{\overline{r}_o}{\alpha \sigma_o^2} - \overline{w}\right) \mathbf{1}_2\right] - \alpha \mathbf{V} \end{aligned} \tag{A-110}$$

and

$$\mathbf{A}_1+\mathbf{D}_{\boldsymbol{\delta}}\mathbf{A}_1\left(\mathbf{I}_4-\boldsymbol{\varrho}\right)+\alpha\left[\mathbf{V}+\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{1}_2'\right]\boldsymbol{\Gamma}\mathbf{A}_1 \qquad \text{(A-111)}$$

$$=\quad \mathbf{1}_2\mathbf{e}_m'-\alpha\left[\mathbf{V}+\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{1}_2'\right]\mathbf{E}_d+\kappa_x\frac{\alpha\sigma_o^2}{\kappa_o^2}\mathbf{1}_2\mathbf{e}_w'.$$

Note that here

$$\mathbf{A}_1+\mathbf{D}_{\boldsymbol{\delta}}\mathbf{A}_1\left(\mathbf{I}-\boldsymbol{\varrho}\right)=\left[\mathbf{E}+\mathbf{D}_{\boldsymbol{\delta}}\mathbf{E}\left(\mathbf{I}-\boldsymbol{\varrho}\right)\right]\circ\mathbf{A}_1,$$

where

$$\mathbf{E}+\mathbf{D}_{\boldsymbol{\delta}}\mathbf{E}\left(\mathbf{I}-\boldsymbol{\varrho}\right)=\begin{bmatrix}1+\frac{1-\rho_m\delta_S}{1-\delta_S} & 1+\frac{1-\rho_{d,S}\delta_S}{1-\delta_S} & 1+\frac{1-\rho_{d,L}\delta_S}{1-\delta_S} & 1+\frac{1-\rho_w\delta_S}{1-\delta_S}\\ 1+\frac{1-\rho_m\delta_L}{1-\delta_L} & 1+\frac{1-\rho_{d,S}\delta_L}{1-\delta_L} & 1+\frac{1-\rho_{d,L}\delta_L}{1-\delta_L} & 1+\frac{1-\rho_w\delta_L}{1-\delta_L}\end{bmatrix}$$

and $\circ$ denotes the Hadamard or element-wide matrix product. Further, we have

$$\begin{aligned}&\alpha\left[\mathbf{V}+\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{1}_2'\right]\boldsymbol{\Gamma}\mathbf{A}_1\\ =\quad&\alpha\begin{bmatrix}V_S+\sigma_o^2\frac{\kappa_x^2}{\kappa_o^2} & C_{S,L}+\sigma_o^2\frac{\kappa_x^2}{\kappa_o^2}\\ C_{S,L}+\sigma_o^2\frac{\kappa_x^2}{\kappa_o^2} & V_L+\sigma_o^2\frac{\kappa_x^2}{\kappa_o^2}\end{bmatrix}\begin{bmatrix}\gamma_SA_m^S & \gamma_SA_{d,S}^S & \gamma_SA_{d,L}^S & \gamma_SA_w^S\\ \gamma_LA_m^L & \gamma_LA_{d,s}^L & \gamma_LA_{d,L}^L & \gamma_LA_w^L\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}&\mathbf{1}\mathbf{e}_m'-\alpha\left[\mathbf{V}+\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{1}_2'\right]\mathbf{E}_d+\kappa_x\frac{\alpha\sigma_o^2}{\kappa_o^2}\mathbf{1}\mathbf{e}_w'\\ =\quad&\begin{bmatrix}1 & -\alpha\left(V_S+\sigma_o^2\frac{\kappa_x^2}{\kappa_o^2}\right) & -\alpha\left(C_{S,L}+\sigma_o^2\frac{\kappa_x^2}{\kappa_o^2}\right) & \kappa_x\frac{\alpha\sigma_o^2}{\kappa_o^2}\\ 1 & -\alpha\left(C_{S,L}+\sigma_o^2\frac{\kappa_x^2}{\kappa_o^2}\right) & -\alpha\left(V_L+\sigma_o^2\frac{\kappa_x^2}{\kappa_o^2}\right) & \kappa_x\frac{\alpha\sigma_o^2}{\kappa_o^2}\end{bmatrix}.\end{aligned}$$

To solve for $\mathbf{A}_1$, we vectorize (A-111) to obtain

$$\begin{aligned}&\left[diag\left(vec\left(\mathbf{E}+\mathbf{D}_{\boldsymbol{\delta}}\mathbf{E}\left(\mathbf{I}_4-\boldsymbol{\varrho}\right)\right)\right)+\mathbf{I}_4\otimes\alpha\left[\mathbf{V}+\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{1}_2'\right]\boldsymbol{\Gamma}\right]vec\left(\mathbf{A}_1\right)\\ &=vec\left[\mathbf{1}_2\mathbf{e}_m'-\alpha\left[\mathbf{V}+\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{1}_2'\right]\mathbf{E}_d+\alpha\frac{1}{\kappa_x}\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{e}_w'\right],\end{aligned}$$

where $\otimes$ denotes the Kronecker product. Solving this equation for $vec\left(\mathbf{A}_1\right)$, we obtain

$$\begin{aligned}vec\left(\mathbf{A}_1\right)=&\left[diag\left(vec\left(\mathbf{E}+\mathbf{D}_{\boldsymbol{\delta}}\mathbf{E}\left(\mathbf{I}_4-\boldsymbol{\varrho}\right)\right)\right)+\mathbf{I}_4\otimes\alpha\left[\mathbf{V}+\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{1}_2'\right]\boldsymbol{\Gamma}\right]^{-1}\\ &\times\ vec\left[\mathbf{1}_2\mathbf{e}_m'-\alpha\left[\mathbf{V}+\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{1}_2'\right]\mathbf{E}_d+\alpha\frac{1}{\kappa_x}\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{e}_w'\right],\end{aligned}$$

where we note that the matrix in square brackets is block-diagonal. Finally, since $\mathbf{V}=\mathbf{D}_{\boldsymbol{\delta}}\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_1'\mathbf{D}_{\boldsymbol{\delta}}'$, we obtain the following fixed-point problem in $\mathbf{A}_1$:

$$\begin{aligned}vec\left(\mathbf{A}_1\right)=&\left[diag\left(vec\left(\mathbf{E}+\mathbf{D}_{\boldsymbol{\delta}}\mathbf{E}\left(\mathbf{I}_4-\boldsymbol{\varrho}\right)\right)\right)+\mathbf{I}_4\otimes\alpha\left[\mathbf{D}_{\boldsymbol{\delta}}\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_1'\mathbf{D}_{\boldsymbol{\delta}}'+\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{1}_2'\right]\boldsymbol{\Gamma}\right]^{-1}\\ &\times\ vec\left[\mathbf{1}\mathbf{e}_m'-\alpha\left[\mathbf{D}_{\boldsymbol{\delta}}\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_1'\mathbf{D}_{\boldsymbol{\delta}}'+\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{1}_2'\right]\mathbf{E}_d+\alpha\frac{1}{\kappa_x}\left(\frac{\kappa_x}{\kappa_o}\right)^2\sigma_o^2\mathbf{1}_2\mathbf{e}_w'\right].\end{aligned}$$

After some algebra, when $\Gamma > 0$ and demand is elastic, the above equation yields

$$\begin{bmatrix} A_m^S \\ A_m^L \end{bmatrix} = \left(\begin{bmatrix} \frac{1-\rho_m \delta_S}{1-\delta_S} & 0 \\ 0 & \frac{1-\rho_m \delta_L}{1-\delta_L} \end{bmatrix} + \alpha \left[\mathbf{V} + \left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 \mathbf{1}_2 \mathbf{1}_2' \right] \boldsymbol{\Gamma} \right)^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$$\begin{bmatrix} A_{d,S}^S \\ A_{d,S}^L \end{bmatrix} = -\left(\begin{bmatrix} \frac{1-\rho_{d,S} \delta_S}{1-\delta_S} & 0 \\ 0 & \frac{1-\rho_{d,S} \delta_L}{1-\delta_L} \end{bmatrix} + \alpha \left[\mathbf{V} + \left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 \mathbf{1}_2 \mathbf{1}_2' \right] \boldsymbol{\Gamma} \right)^{-1} \begin{bmatrix} \alpha \left[\left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 + V_S \right] \\ \alpha \left[\left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 + C_{S,L} \right] \end{bmatrix},$$

$$\begin{bmatrix} A_{d,L}^S \\ A_{d,L}^L \end{bmatrix} = -\left(\begin{bmatrix} \frac{1-\rho_{d,L} \delta_S}{1-\delta_S} & 0 \\ 0 & \frac{1-\rho_{d,L} \delta_L}{1-\delta_L} \end{bmatrix} + \alpha \left[\mathbf{V} + \left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 \mathbf{1}_2 \mathbf{1}_2' \right] \boldsymbol{\Gamma} \right)^{-1} \begin{bmatrix} \alpha \left[\left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 + C_{S,L} \right] \\ \alpha \left[\left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 + V_L \right] \end{bmatrix},$$

$$\begin{bmatrix} A_w^S \\ A_w^L \end{bmatrix} = \alpha \frac{1}{\kappa_x} \left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 \left(\begin{bmatrix} \frac{1-\rho_w \delta_S}{1-\delta_S} & 0 \\ 0 & \frac{1-\rho_w \delta_L}{1-\delta_L} \end{bmatrix} + \alpha \left[\mathbf{V} + \left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 \mathbf{1}_2 \mathbf{1}_2' \right] \boldsymbol{\Gamma} \right)^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

If $\gamma_S = \gamma_L = 0$, these equations simplify to

$$\begin{bmatrix} A_m^S & A_{d,S}^S & A_{d,L}^S & A_w^S \\ A_m^L & A_{d,S}^L & A_{d,L}^L & A_w^L \end{bmatrix} = \begin{bmatrix} \frac{1-\delta_S}{1-\delta_S \rho_m} & \frac{1-\delta_L}{1-\delta_L \rho_m} \\ -\frac{1-\delta_S}{1-\delta_S \rho_{d,S}} \alpha \left[\left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 + V_S \right] & -\frac{1-\delta_L}{1-\delta_L \rho_{d,S}} \alpha \left[\left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 + C_{S,L} \right] \\ -\frac{1-\delta_S}{1-\delta_S \rho_{d,L}} \alpha \left[\left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 + C_{S,L} \right] & -\frac{1-\delta_L}{1-\delta_L \rho_{d,L}} \alpha \left[\left(\frac{\kappa_x}{\kappa_o} \right)^2 \sigma_o^2 + V_L \right] \\ \frac{1-\delta_S}{1-\delta_S \rho_w} \kappa_x \frac{\alpha \sigma_o^2}{\kappa_o^2} & \frac{1-\delta_L}{1-\delta_L \rho_w} \kappa_x \frac{\alpha \sigma_o^2}{\kappa_o^2} \end{bmatrix}^\top; \tag{A-112}$$

which is the obvious generalization of our baseline model when $\gamma = 0$.

The following Proposition summarizes some results on the swap spread coefficients for short- and long-dated swaps:

Proposition 16 *If an equilibrium exists, we have A_m^i,,$A_w^i > 0$ and $A_{d,S}^i$,$A_{d,L}^i < 0$ for $i = S, L$. Moreover, if $\gamma_L = \gamma_S = 0$, shocks to the short-rate different and intermediary supply have a larger impact on short-maturity spreads than on long-maturity swaps: we have $A_m^S > A_m^L > 0$ and $A_w^S > A_w^L > 0$. However, demand shocks can have a hump-shaped effect on the spread curve: depending on the parameters, we can have either $0 > A_{d,j}^i > A_{d,j}^{-i}$ or $0 > A_{d,j}^{-i} > A_{d,j}^i$ for all $i, j = S, L$.*

Proof. The first sentence follows directly from (A-112). The second sentence follows from the fact that $\frac{1-\delta_L}{1-\delta_L \rho} < \frac{1-\delta_S}{1-\delta_S \rho}$ for all $\rho < 1$ since $\delta_L > \delta_S$. It then follows immediately that $A_m^S > A_m^L > 0$ and $A_w^S > A_w^L > 0$.

Finally, define $Corr_{S,L} = C_{S,L} / \sqrt{V_L V_S}$ and assume that $V_L > V_S > 0$ and $Corr_{S,L} \in (0, 1)$ in the stable equilibrium, implying that $C_{S,L} = Corr_{S,L} \sqrt{V_L} \sqrt{V_S} < V_L$. If shocks to the demand for long-dated swaps are highly persistent, these shocks will have a larger impact on longer dated swaps. Formally, if $\rho_{d,L} \to 1$, we are going to have $0 > A_{d,L}^S > A_{d,L}^L$ However, when long-dated demand is less persistent, shocks to long-dated demand can have a hump-shaped effect on the spread curve—i.e., we can have $0 > A_{d,L}^L > A_{d,L}^S$.

At the same time, $C_{S,L}$ can exceed V_S if $Corr_{S,L}$ is large enough or can be less than V_S if $Corr_{S,L}$ is low enough. If $V_S > C_{S,L}$, then we must have $0 > A_{d,S}^L > A_{d,S}^S$—i.e., short-dated demand shocks will have a bigger impact on short-dated spreads. However, if $C_{S,L} > V_S$, we can either have $0 > A_{d,S}^L > A_{d,S}^S$ or $0 > A_{d,S}^S > A_{d,S}^L$. And the greater is the difference between $C_{S,L}$ and V_S, the more likely we are to have $0 > A_{d,S}^S > A_{d,S}^L$. ■

B.5.2 Empirical predictions

Empirically, we would then estimate these latent factors using a VAR that imposes a combination of zero, sign, and monotonicity restrictions. We rewrite our model, using short-dated spread as m_t and swap prices and quantities at two points on the curve, short (S) and long (L):

$$\overbrace{\begin{bmatrix} m_t \\ s_t^S \\ s_t^L \\ d_t^S \\ d_t^L \end{bmatrix}}^{\mathbf{y}_t} = \overbrace{\begin{bmatrix} \bar{m} \\ A_0^S \\ A_0^L \\ \left(\bar{d}_S + \gamma_S A_0^S\right) \\ \left(\bar{d}_L + \gamma_L A_0^L\right) \end{bmatrix}}^{\mathbf{a}} + \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ A_m^S & A_{d,S}^S & A_{d,L}^S & A_w^S & . \\ A_m^L & A_{d,S}^L & A_{d,L}^L & A_w^L & . \\ \gamma_S A_m^S & 1 + \gamma_S A_{d,S}^S & \gamma_S A_{d,L}^S & \gamma_S A_w^S & . \\ \gamma_L A_m^L & \gamma_L A_{d,S}^L & 1 + \gamma_L A_{d,L}^L & \gamma_L A_w^L & . \end{bmatrix}}^{\mathbf{A}} \overbrace{\begin{bmatrix} z_t^m \\ z_t^{d,S} \\ z_t^{d,L} \\ z_t^w \\ z_t^U \end{bmatrix}}^{\mathbf{z}_t}, \tag{A-113}$$

where z_t^U is a non-identified shock orthogonal to the other shocks. Thus, z_t^U reflects unmodeled factors that may impact equilibrium swap spreads and quantities. Empirically we think about S as being 5 years and L as 30 years.

Based on our theoretical results, assuming that the γs are not too large, the above model implies the following sign and monotonicity restrictions on swap spreads and positions in (A-113):

- Monotonicity of spreads in the short-date differential:

$$0 < A_m^L < A_m^S < 1;$$

- Monotonicity of spreads in intermediary wealth:

$$0 < A_w^L < A_w^S;$$

- Negative loading of spreads on the short-dated demand shifter:

$$A_{d,S}^S < 0 \text{ and } A_{d,S}^L < 0;$$

- Negative loading of spreads on the long-dated demand shifter:

$$A_{d,L}^S < 0 \text{ and } A_{d,L}^L < 0;$$

- Positive impact of the short-date differential on quantities:

$$\gamma_S A_m^S > 0 \text{ and } \gamma_L A_m^L > 0;$$

- Positive impact of intermediary wealth on quantities:

$$\gamma_S A_w^S > 0 \text{ and } \gamma_L A_w^L > 0;$$

- Monotonic impact of the short-dated demand factor on quantities:

$$1 + \gamma_S A_{d,S}^S > \gamma_L A_{d,S}^L;$$

- Monotonic impact of the long-dated demand on quantities:

$$1 + \gamma_L A_{d,L}^L > \gamma_S A_{d,L}^S.$$

Alternatively, we can rewrite (A-113) as replacing the short- and long-maturity equilibrium positions with the aggregate equilibrium swap position and the difference of long- vs short-maturity holdings (and also ignoring the interest rate differential and the impact of the fundamental shocks) to obtain:

$$\overbrace{\begin{bmatrix} s_t^S \\ s_t^L \\ d_t^S + d_t^L \\ d_t^L - d_t^S \end{bmatrix}}^{\mathbf{y}_t} = \mathbf{a} + \overbrace{\begin{bmatrix} A_{d,S}^S & A_{d,L}^S & A_w^S & . \\ A_{d,S}^L & A_{d,L}^L & A_w^L & . \\ 1 + \gamma_S A_{d,S}^S + \gamma_L A_{d,S}^L & 1 + \gamma_S A_{d,L}^S + \gamma_L A_{d,L}^L & \gamma_S A_w^S + \gamma_L A_w^L & . \\ \gamma_L A_{d,S}^L - 1 - \gamma_S A_{d,S}^S & 1 + \gamma_L A_{d,L}^L - \gamma_S A_{d,L}^S & \gamma_L A_w^L - \gamma_S A_w^S & . \end{bmatrix}}^{\mathbf{A}} \overbrace{\begin{bmatrix} z_t^{d,S} \\ z_t^{d,L} \\ z_t^w \\ z_t^U \end{bmatrix}}^{\mathbf{z}_t}, \tag{A-114}$$

with $\mathbf{a} = [A_0^S, A_0^L, (\bar{d}_S + \gamma_S A_0^S) + (\bar{d}_L + \gamma_L A_0^L), (\bar{d}_L + \gamma_L A_0^L) - (\bar{d}_S + \gamma_S A_0^S)]'$. From here, we obtain the following model-implied sign restrictions:

- Shocks to the demand for short-term swaps lower the spreads on both short- and long-term swap spreads

$$A_{d,S}^S < 0 \text{ and } A_{d,S}^L < 0,$$

while increasing the total equilibrium quantity

$$1 + \gamma_S A_{d,S}^S + \gamma_L A_{d,S}^L > 0,$$

and decreasing the difference in the long- and short-term quantities,

$$\gamma_L A_{d,S}^L - 1 - \gamma_S A_{d,S}^S < 0;$$

- Shocks to the demand for long-term swaps also reduce both short- and long-term swap spreads,

$$A_{d,L}^S < 0 \text{ and } A_{d,L}^L < 0,$$

while increasing the overall dealer net bond position,

$$1 + \gamma_S A^S_{d,L} + \gamma_L A^L_{d,L} > 0,$$

and increasing the difference in the long- and short-term quantities,

$$1 + \gamma_L A^L_{d,L} - \gamma_S A^S_{d,L} > 0;$$

- Finally, intermediary wealth shocks increase both short- and long-term swap spreads,

  $$A^S_w > 0 \text{ and } A^L_w > 0$$

  increase the total equilibrium quantity,

  $$\gamma_S A^S_w + \gamma_L A^L_w > 0,$$

  but has an ambiguous impact on the difference in the long- and short-term quantities. (However, the impact on the difference between long- and short-term quantities becomes unambiguous if $\gamma_L = \gamma_S = \gamma$. In this case, we have $\gamma_L A^L_w - \gamma_S A^S_w = \gamma \left(A^L_w - A^S_w\right) < 0$.)

In summary, the model implies the following set of sign restrictions

$$\overbrace{\begin{bmatrix} \xi^{s^S} \\ \xi^{s^L} \\ \xi^{d^S+d^L} \\ \xi^{d^L-d^S} \end{bmatrix}}^{\boldsymbol{\xi}_t} = \overbrace{\begin{bmatrix} - & - & + & . \\ - & - & + & . \\ + & + & + & . \\ - & + & . & . \end{bmatrix}}^{\mathbf{A}_3} \overbrace{\begin{bmatrix} \varepsilon^{d,S}_t \\ \varepsilon^{d,L}_t \\ \varepsilon^{w}_t \\ \varepsilon^{U}_t \end{bmatrix}}^{\boldsymbol{\varepsilon}_t}. \quad \text{(A-115)}$$

Appendix C.2 contains the results of performing a sign-restricted VAR on our data based on exactly these restrictions; see (A-116).

C Additional empirical results

C.1 Additional term structure implications

Setting aside the short-rate differential ($m_t = 0$) and assuming that end-user demand is inelastic ($\gamma = 0$), our baseline model implies that the term structure of swap spreads depends on just two factors: the level of end-user demand for long-term swaps as a whole (d_t) and the amount of intermediary capital (w_t). Specifically, n-period swap spreads are

$$s^{(n)}_t = A^{(n)}_0 + A^{(n)}_d \left(d_t - \overline{d}\right) + A^{(n)}_w \left(w_t - \overline{w}\right),$$

where $A^{(n)}_0$ is negative and decreasing in n, $A^{(n)}_w$ is positive and decreasing in n, and $A^{(n)}_d$ is negative and is either decreasing in n or is a U-shaped function of n.

Thus, our model suggests that, in absolute terms, supply shocks should have a larger impact on short-dated swaps while demand shocks should have their greatest impact on intermediate- or longer-dated swaps. In Tables D-1 and D-2, we test this additional prediction by regressing the 3-month change in swap spreads at various maturities n on various proxies for the change in demand and supply:

$$\Delta_{13} s^{(n)}_t = \alpha^{(n)} + \beta^{(n)} \cdot \Delta X_t + \Delta_{13} \varepsilon^{(n)}_t.$$

We also regress changes in the slope of the swap spread curve from year m to n, namely, $\Delta_{13}(s^{(n)}_t - s^{(m)}_t)$, on changes in these same proxies:

$$\Delta_{13}(s^{(n)}_t - s^{(m)}_t) = \alpha^{(n)} + \beta^{(n)} \cdot \Delta X_t + \Delta_{13} \varepsilon^{(n)}_t.$$

As shown in Table D-1 below, the coefficients for our demand proxies are generally consistent with this prediction. Specifically, changes in demand generally have their largest impact on longer-dated spreads and, thus, alter differences between longer- and shorter-dated swap spreads in the way predicted by our theory. However, as shown in Table D-2, the results for our supply proxies are less clear cut. Specifically, while changes in the Adrian et al. (2014) broker-dealer leverage factor, $\Delta_{13} \ln (BD\text{-}Lev_t)$, have a larger impact on shorter-dated spreads as predicted, the results hedge fund returns ($\Delta_{13} \ln (HFR_t)$) tend to go in the wrong way. Overall, we view this exercise as a partial win for our theory, but this is clearly an area that should be explored in future research.

C.2 Additional VARs

To check that our identification of demand and supply shocks is not sensitive to the inclusion of the LIBOR-repo spread (m_t) as an additional variable in the VAR, we consider the following tri-variate specification:

$$\begin{bmatrix} m_t \\ s_t^{(30)} \\ \hat{d}_t \end{bmatrix} = \mathbf{c} + \sum_{l=1}^{L} \mathbf{C}_l \begin{bmatrix} m_{t-l} \\ s_{t-l}^{(30)} \\ \hat{d}_{t-l} \end{bmatrix} + \xi_t.$$

Following Proposition 4, we identify the structural demand, supply, and LIBOR-repo shocks by imposing a combination of sign and zero restrictions. Specifically, in addition to the structural shock orthogonality and the sign restrictions that we imposed in our baseline VAR, we also assume that the LIBOR-repo spread does not respond on impact to demand and supply shocks, and that the on-impact responses of the LIBOR-repo spread, of the 30-year swap spread and of end-users' demand to the LIBOR-repo shock have the same sign. Thus, structural shocks ε_t are related to reduced-form VAR residuals ξ_t by the mapping

$$\overbrace{\begin{bmatrix} \xi_t^m \\ \xi_t^{s^{(30)}} \\ \xi_t^{\hat{d}} \end{bmatrix}}^{\boldsymbol{\xi}_t} = \overbrace{\begin{bmatrix} + & 0 & 0 \\ + & - & + \\ + & + & + \end{bmatrix}}^{\mathbf{A}} \overbrace{\begin{bmatrix} \varepsilon_t^m \\ \varepsilon_t^d \\ \varepsilon_t^w \end{bmatrix}}^{\boldsymbol{\varepsilon}_t}.$$

We estimate the structural VAR using the sign and zero restriction approach of Arias et al. (2018) with lag length set to $L = 4$. The historical decomposition implied by the estimated VAR and shown on Figure D-1 suggests that LIBOR-repo shocks contribute very little to the swap spread variation. This is confirmed by the forecast error variance decomposition: LIBOR-repo shocks account for approximately 2% of the swap spread variance in the long run.

To study maturity-specific end-user demand to receive fixed rate, we consider the following VAR specification:

$$\begin{bmatrix} s_t^{(5)} \\ s_t^{(30)} \\ \hat{d}_t \\ \hat{d}_t^L - \hat{d}_t^S \end{bmatrix} = \mathbf{c} + \sum_{l=1}^{L} \mathbf{C}_l \begin{bmatrix} s_{t-l}^{(5)} \\ s_{t-l}^{(30)} \\ \hat{d}_{t-l} \\ \hat{d}_{t-l}^L - \hat{d}_{t-l}^S \end{bmatrix} + \boldsymbol{\xi}_t.$$

We identify the structural short-maturity demand, long-maturity demand, and supply shocks by imposing a combination of sign restrictions. First, we assume that positive demand shocks (both short-maturity and long-maturity) make swap spreads (both $s_t^{(5)}$ and $s_t^{(30)}$) more negative and increase the scale of primary dealers' total position $\hat{d}_t = \hat{d}_t^L + \hat{d}_t^S$. Moreover, we assume that shocks to short-maturity demand have a larger effect on intermediaries' net position in Treasuries with maturities of less than 6 years (*PD-UST-Net*$_t^{<y} = \hat{d}_t^S$) relative to their net position in Treasuries with maturities of more than 6 years (*PD-UST-Net*$_t^{>6} = \hat{d}_t^L$) and vice versa. Thus, a positive short-maturity demand shock on impact decreases $\hat{d}_t^L - \hat{d}_t^S$, while a positive long-maturity demand shock increases it. Finally, we assume that a positive shock to intermediary supply makes swap spreads (both $s_t^{(5)}$ and $s_t^{(30)}$) less negative, while increasing intermediaries' overall short position $\hat{d}_t$. The fourth shock in the VAR is not identified. Thus, structural shocks $\boldsymbol{\varepsilon}_t$ are related to reduced-form VAR residuals $\boldsymbol{\xi}_t$ by the mapping

$$\overbrace{\begin{bmatrix} \xi^{s^{(5)}} \\ \xi^{s^{(30)}} \\ \xi^{\hat{d}} \\ \xi^{\hat{d}^L - \hat{d}^S} \end{bmatrix}}^{\boldsymbol{\xi}_t} = \overbrace{\begin{bmatrix} - & - & + & . \\ - & - & + & . \\ + & + & + & . \\ - & + & . & . \end{bmatrix}}^{\mathbf{A}_3} \overbrace{\begin{bmatrix} \varepsilon_t^{d,S} \\ \varepsilon_t^{d,L} \\ \varepsilon_t^w \\ \varepsilon_t^U \end{bmatrix}}^{\boldsymbol{\varepsilon}_t} \quad \text{(A-116)}$$

We estimate the structural VAR with the pure sign restrictions approach of Uhlig (2005) with lag length $L = 2$.

C.3 Demand and supply as return predictors

Proposition 3 states that higher (lower) end-user demand for receiving fixed and lower (higher) intermediary supply of pay-fixed swaps are associated with higher (lower) expected returns on the short position in the swap arbitrage trade. To gauge the respective effects of end-user demand and intermediary supply, we regress swap arbitrage returns on our demand and supply factors:

$$r_{t\to t+h}^{s(30)} = \alpha + \beta_1 \cdot \mathit{Demand}_t + \beta_2 \cdot \mathit{Supply}_t + \epsilon_{t\to t+h} \quad \text{(A-117)}$$

As reported in Panel A of Table D-4, the demand factor strongly predicts returns to the 30-year swap spread trade at both 3-month and 12-month horizons. As reported in Panel B, demand also predicts returns to 10-year swap spread trade, albeit with lower statistical significance. The negative sign of the estimated β_1 coefficient implies that a higher level of the demand factor corresponds to higher expected returns on the short position in the swap arbitrage trade, in line with our model's prediction. The R^2 of the predictive regression for 3-month returns on the 30-year swap spread trade reaches 13.0%, similar to that in Table 1.

By contrast, the supply factor is not statistically significant in the predictive regressions. Our theory helps explain the fact that supply is not a strong predictor of returns to swap arbitrage at short-horizons even though it makes an important contribution to variation in the level of swap spreads. In contrast to demand shocks, supply shocks move the shadow value of intermediary capital and compensation for bearing swap spread risk in opposite directions. This mitigates the effect of supply shifts on expected returns at short horizons. At the same time, as discussed above, the effect of supply shocks on expected returns is persistent. Thus, even though they have smaller effects on expected returns at short-horizons, supply shifts have a substantial effect on the average expected return over the life of the swap which determines the level of swap spreads.[54]

[54]Our demand and supply factors are a rotation of $s_t^{(30)}$ and $\textit{PD-Net-UST}_t$. Thus, the predictive power of these two sets of factors is the same by construction. Interestingly, as shown in columns (2) and (4), $\textit{PD-Net-UST}_t$ contains incremental information about expected returns beyond that contained in $s_t^{(30)}$. Why is this the case? First, fluctuations in the short-rate differential (m_t) mean that long-term LIBOR swap spreads are not a sufficient statistic for the expected long-run returns on the swap spread trade. Second, even if $m_t \equiv 0$, $\textit{PD-Net-UST}_t$ will contain incremental information about 1-period-ahead returns so long as demand and supply shocks have different persistences. If supply shocks are more persistent, we would expect $\textit{PD-Net-UST}_t$ to attract a negative coefficient in this bivariate regression.

D Additional tables and figures

Table D-1: **Demand shifters and the shape of the swap spread curve:** Panel A reports the coefficients from regressions of the 3-month change in swap spreads at various maturities $n = 2$, 5, 10, and 30 on various proxies for the change in demand:

$$\Delta_{13} s_t^{(n)} = \alpha^{(n)} + \beta^{(n)} \cdot \Delta X_t^d + \Delta_{13}\varepsilon_t^{(n)}.$$

In Panel B, we regress 3-month changes in the slope of the swap spread curve from year m to n, namely, $\Delta_{13}(s_t^{(n)} - s_t^{(m)})$, on changes in these same proxies:

$$\Delta_{13}(s_t^{(n)} - s_t^{(m)}) = \alpha^{(n)} + \beta^{(n)} \cdot \Delta X_t^d + \Delta_{13}\varepsilon_t^{(n)}.$$

The demand proxies (X_t^d) include modified duration of the Barclays U.S. MBS index (*MBS-Duration*$_t$), the Klingler and Sundaresan (2019) pension underfunding factor (*Pension-UFR*$_t$), and corporate bond issuance (*Corp-Issuance*$_t$). Data are weekly and run from January 2009 to June 2018. Newey and West (1987) t-statistics are reported in parentheses and are computed with 20 lags. We compute the associated p-values using the fixed-b asymptotic theory of Kiefer and Vogelsang (2005). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectivelely.

	$\Delta_{13} s_t^{(2)}$	$\Delta_{13} s_t^{(5)}$	$\Delta_{13} s_t^{(10)}$	$\Delta_{13} s_t^{(30)}$
	(1)	(2)	(3)	(4)
Panel A: Spreads				
Δ_{13}*MBS-Duration*$_t$	0.533	0.144	2.357***	3.728***
	(0.45)	(0.19)	(2.99)	(5.01)
Δ_{13}*Pension-UFR*$_t$	2.263***	1.481***	0.833	-1.056
	(3.74)	(3.58)	(1.69)	(1.46)
Δ_{13}*Corp-Issuance*$_t$	0.023	0.043	-0.009	-0.114**
	(0.50)	(1.22)	(0.25)	(2.36)
Adjusted R^2	0.136	0.088	0.085	0.196
N	482	482	482	482
	$\Delta_{13}(s_t^{(5)} - s_t^{(2)})$	$\Delta_{13}(s_t^{(10)} - s_t^{(5)})$	$\Delta_{13}(s_t^{(30)} - s_t^{(10)})$	$\Delta_{13}(s_t^{(30)} - s_t^{(2)})$
Panel B: Slopes				
Δ_{13}*MBS-Duration*$_t$	-0.389	2.213***	1.371**	3.195***
	(0.71)	(3.13)	(2.08)	(2.68)
Δ_{13}*Pension-UFR*$_t$	-0.782*	-0.648	-1.889***	-3.319***
	(1.90)	(1.49)	(4.21)	(5.29)
Δ_{13}*Corp-Issuance*$_t$	0.020	-0.052**	-0.105**	-0.137**
	(0.65)	(2.29)	(2.32)	(2.65)
Adjusted R^2	0.066	0.164	0.242	0.289
N	482	482	482	482

Table D-2: **Supply shifters and the shape of the swap spread curve:** Panel A reports the coefficients from regressions of the 3-month change in swap spreads at various maturities n = 2, 5, 10, and 30 on various proxies for the change in supply:

$$\Delta_{13} s_t^{(n)} = \alpha^{(n)} + \beta^{(n)} \cdot \Delta X_t^s + \Delta_{13} \varepsilon_t^{(n)}.$$

In Panel B, we regress 3-month changes in the slope of the swap spread curve from year m to n, namely, $\Delta_{13}(s_t^{(n)} - s_t^{(m)})$, on changes in these same proxies:

$$\Delta_{13}(s_t^{(n)} - s_t^{(m)}) = \alpha^{(n)} + \beta^{(n)} \cdot \Delta X_t^d + \Delta_{13} \varepsilon_t^{(n)}.$$

The supply proxies (X_t^s) include the broker-dealer leverage factor from Adrian et al. (2014) (*BD-Lev*$_t$) and the logarithm of the HFRX Global Hedge Fund Index ($\ln(HFR_t)$). Data are weekly and run from January 2009 to June 2018. Newey and West (1987) t-statistics are reported in parentheses and are computed with 20 lags. We compute the associated p-values using the fixed-b asymptotic theory of Kiefer and Vogelsang (2005). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	$\Delta_{13} s_t^{(2)}$	$\Delta_{13} s_t^{(5)}$	$\Delta_{13} s_t^{(10)}$	$\Delta_{13} s_t^{(30)}$
	(1)	(2)	(3)	(4)
Panel A: Spreads				
$\Delta_{13}\ln(BD\text{-}Lev_t)$	70.492	78.696**	8.225	-0.219
	(1.65)	(2.67)	(0.32)	(0.01)
$\Delta_{13}\ln(HFR_t)$	-0.789	-0.441	0.070	1.097**
	(1.39)	(1.04)	(0.15)	(2.24)
Adjusted R^2	0.105	0.125	-0.003	0.080
N	482	482	482	482
	$\Delta_{13}(s_t^{(5)} - s_t^{(2)})$	$\Delta_{13}(s_t^{(10)} - s_t^{(5)})$	$\Delta_{13}(s_t^{(30)} - s_t^{(10)})$	$\Delta_{13}(s_t^{(30)} - s_t^{(2)})$
Panel B: Slopes				
$\Delta_{13}\ln(BD\text{-}Lev_t)$	8.205	-70.472**	-8.444	-70.711
	(0.35)	(2.71)	(0.33)	(1.35)
$\Delta_{13}\ln(HFR_t)$	0.349	0.511**	1.026***	1.886***
	(1.48)	(2.27)	(2.82)	(3.97)
Adjusted R^2	0.020	0.210	0.120	0.206
N	482	482	482	482

Table D-3: **Short-term and long-term demand factors, 2009m1 to 2018m6:** This table reports the slope coefficients from regressions of, respectively, the 5-year swap spread ($s_t^{(5)}$), the 30-year swap spread ($s_t^{(30)}$), the short-term demand factor ($ST\text{-}Demand_t$), and the long-term demand factor ($LT\text{-}Demand_t$) on the modified duration of the Barclays U.S. MBS index ($MBS\text{-}Duration_t$), the Klingler and Sundaresan (2019) pension underfunding factor ($Pension\text{-}UFR_t$), and corporate bond issuance ($Corp\text{-}Issuance_t$). Variables are in levels in Panel A and in 13-week changes in Panel B. Data are weekly and run from January 2009 to June 2018. Newey and West (1987) t-statistics are reported in parentheses and are computed with 78 lags in Panel A and with 20 lags in Panel B. We compute the associated p-values using the fixed-b asymptotic theory of Kiefer and Vogelsang (2005). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	$s_t^{(5)}$	$s_t^{(30)}$	$ST\text{-}Demand_t$	$LT\text{-}Demand_t$
	(1)	(2)	(3)	(4)
Panel A: Levels				
$MBS\text{-}Duration_t$	-4.747**	5.557*	-0.064	-1.622
	(3.20)	(2.24)	(0.09)	(1.63)
$Pension\text{-}UFR_t$	0.314	-1.991	-0.230	1.117*
	(0.16)	(1.30)	(0.45)	(2.30)
$Corp\text{-}Issuance_t$	-0.248**	-0.250	0.131**	0.064
	(2.52)	(1.88)	(2.49)	(1.33)
Adjusted R^2	0.284	0.283	0.233	0.353
N	495	495	495	495
Panel B: 3-month changes				
$\Delta_{13}MBS\text{-}Duration_t$	0.144	3.728***	-0.464*	-1.596***
	(0.19)	(5.01)	(1.77)	(5.75)
$\Delta_{13}Pension\text{-}UFR_t$	1.480***	-1.057	-0.011	0.806***
	(3.58)	(1.46)	(0.06)	(3.59)
$\Delta_{13}Corp\text{-}Issuance_t$	0.043	-0.114**	0.025*	0.029
	(1.22)	(2.36)	(1.83)	(1.52)
Adjusted R^2	0.088	0.196	0.057	0.349
N	482	482	482	482

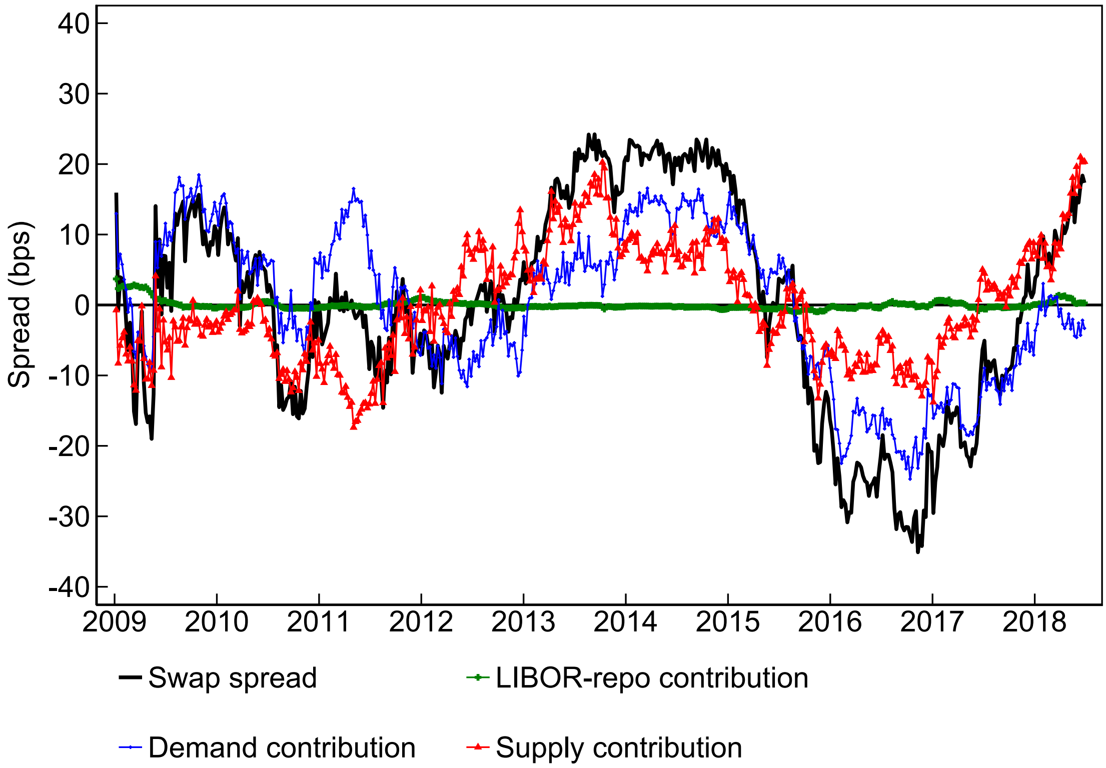
Table D-4: **Forecasting swap spread trade returns using the demand and supply factors:** This table reports the slope coefficients from regressions of the form

$$r^{s(n)}_{t\to t+h} = \alpha + \beta_1 \cdot \mathit{Demand}_t + \beta_2 \cdot \mathit{Supply}_t \ + \epsilon_{t\to t+h}$$

for $n = 10$ and 30 years. In words, we regress the returns to the n-year swap spread trade on the estimated demand factor (Demand_t) and the supply factor (Supply_t). We also report the coefficients from the corresponding regressions on $\mathit{PD\text{-}UST\text{-}Net}_t$ and $s^{(30)}_t$ which have the same R^2 by construction. Returns are for the 30-year swap spread trade in Panel A and for the 10-year swap spread trade in Panel B. The holding period is equal to $h = 13$ weeks (3 months) in columns 1 and 2 and to $h = 52$ weeks (12 months) in columns 3 and 4. Data are weekly and run from January 2009 to June 2018. Newey and West (1987) t-statistics computed with 20 lags for 3-month return regressions and 78 lags for 12-month return regressions are reported in parentheses. We compute the associated p-values using the fixed-b asymptotic theory of Kiefer and Vogelsang (2005). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	h = 3-month returns		h = 12-month returns	
	(1)	(2)	(3)	(4)
Panel A: Returns on the 30-year swap spread trade, $r^{s(30)}_{t\to t+h}$				
Demand_t	-13.752***		-55.944***	
	(3.51)		(6.06)	
Supply_t	-1.646		9.666	
	(0.37)		(0.71)	
$\mathit{PD\text{-}UST\text{-}Net}_t$		-1.096***		-3.304**
		(2.75)		(2.86)
$s^{(30)}_t$		2.677*		14.531***
		(1.92)		(3.93)
Adjusted R^2	0.130	0.130	0.508	0.508
N	495	495	495	495
Panel B: Returns on the 10-year swap spread trade, $r^{s(10)}_{t\to t+h}$				
Demand_t	-2.684*		-11.707***	
	(1.93)		(4.41)	
Supply_t	2.063		3.356	
	(1.03)		(0.85)	
$\mathit{PD\text{-}UST\text{-}Net}_t$		-0.046		-0.597
		(0.23)		(1.44)
$s^{(30)}_t$		1.053**		3.337***
		(2.46)		(4.56)
Adjusted R^2	0.069	0.069	0.317	0.317
N	495	495	495	495

Figure D-1: **Swap spread historical decomposition:** This figure shows how LIBOR-repo, end-user demand, and intermediary supply contribute to the time-series variation in 30-year LIBOR swap spread. The underlying data are weekly and run from January 2009 to June 2018. Shocks are identified using the structural VAR described in Appendix C.2. Formally, we plot:



$$s_t^{(30)} - \mathrm{E}[s_t^{(30)}] = \overbrace{\hat{A}_m \hat{z}_t^m}^{\text{Short-rate contribution}} + \overbrace{\hat{A}_d \hat{z}_t^d}^{\text{Demand contribution}} + \overbrace{\hat{A}_w \hat{z}_t^w}^{\text{Supply contribution}}$$

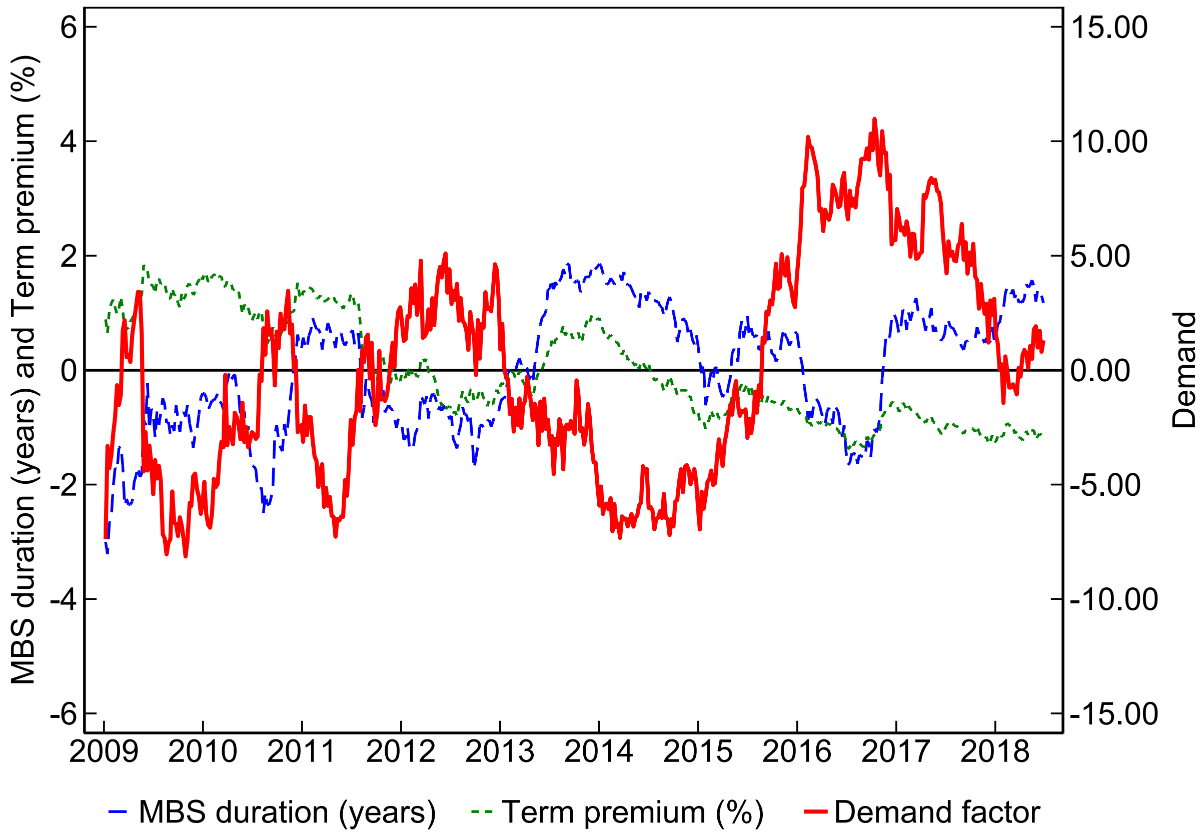


Figure D-2: **Swap demand and MBS duration:** This figure shows the the modified duration of the Barclays U.S. MBS index (MBS duration), the Adrian et al. (2013) term premium (Term premium) and the swap demand factor (Demand factor). Data are weekly and run from January 2009 to June 2018.