A Model of Optimal QE and QT with Segmented Bond Market and Liquidity Regulations

Yinjie Yu

October 23, 2023

Abstract

This paper analyzes optimal asset-purchase and its exit policies in a macroeconomic model with banks facing liquidity regulations (Liquidity Stress Testing and Leverage Coverage Ratio) and balance sheet cost. QE is effective in response to adverse productivity, liquidity, and bank net worth shocks. It works by inflating the bond price, restoring bank net worth, but most importantly through liquidity provision, and thus is more powerful when the LCR constraint binds. The optimal exiting strategy is gradual whenever LCR constraint binds. However, a fast QT is costless and even expansionary when liquidity is abundant because it drives the banks to substitute bonds for working capital.

1 Introduction

Since the global financial crisis, the Fed has implemented various unconventional monetary policies in coordination with the fiscal authority to stabilize the financial market from the turmoil caused by major intermediation failures such as the bankruptcy of Lehman Brothers. As a result, the central bank balance sheet expanded from $900 billion pre-GFC to over $4.5 trillion. The total supply of bank reserves and treasuries surged, and the banking system has transformed into an ample reserve regime.

In contrast to the prompt action of quantitative easing, the central bank hesitates to unwind its expanded balance sheet. Given the scarcity of knowledge of how quantitative monetary policies work in an ample reserve regime, the monetary authority was more cautious about taking action.

After staying neutral for three years, the Fed started an endeavor to shrink its balance sheet from late 2017 to 2019, which was terminated due to the repo market spike in Sept 2019. Before the Fed could resume the unwinding process, the COVID-19 pandemic hit and called for another fiscal and monetary stimulus round that reversed previous efforts. Recently, the Fed announced a new plan of quantitative tightening to lower its position on a wide range of nominal assets, including agency debt, agency MBS, and Treasuries, at a pace twice as fast as the previous round.[1] Up to the point I write the paper, we have not seen significant adverse effects except for the latest local banking crisis featuring the collapse of SVB that was mainly due to an interest rate hike in tandem with QT. However, SVB fell due to overly relying on treasury bonds in its portfolio, which is as sensitive to interest rate hikes as QT, warns that the unwinding process might negatively affect the banking system and trigger more considerable economic unrest.

Debates on whether and how the Fed should shrink its balance sheet remain high policy interest. Greenwood et al. [2016] and Lopez-Salido and Vissing-Jorgensen [2023] argue that the banking system under current regulation and monetary policy environment needs more reserves to maintain financial stability. Figure 1 shows that reserves remained persistently high during a prolonged period after QE ended in 2014 and only started to drop faster after QT. However, with only a 30% reduction of reserves, whose total supply was still far higher than the pre-GFC level, the process was stopped by a disturbance in the repo market. The Fed nevertheless regards an expanded balance sheet as costly and is determined to initiate the tightening process. In a statement in Nov 2018, the FOMC mentioned that it faces challenges in *"precisely determining the quantity of reserves necessary"* in an ample reserve regime and bearing the *"large ongoing interest expenses associated with the remuneration of reserves"*. The Fed also wants to improve the effectiveness of such policy in the future by convincing the public that QE serves as a particular policy instrument in response to financial stress and that the Fed can control its balance sheet size.

With a phenomenally large balance sheet and an even higher reserve level, whether the ongoing unwinding process might again induce undesirable consequences and how it should be optimally designed remains an open question. Should the central bank shrink its balance sheet immediately

[1] On May 4th, 2022, the FOMC announced its *Plans for Reducing the Size of the Federal Reserve's Balance Sheet* in which it revealed that the Fed will reduce its holding on Treasury bonds at a pace of $30 billion per month over the summer and $60 billion per month from now on. Later, they accelerated it to $90 billion monthly.
https://www.federalreserve.gov/newsevents/pressreleases/monetary20220504b.htm

after stabilizing the economy? Should it adopt a more gradual regime or even never try to unwind QE, as a large balance sheet does not prove to have significant costs but benefits to the financial system? How much should the central bank respond to shocks of a different nature with QE policies in the first place? I provide my views on these questions through the lens of a macroeconomic model with banks.

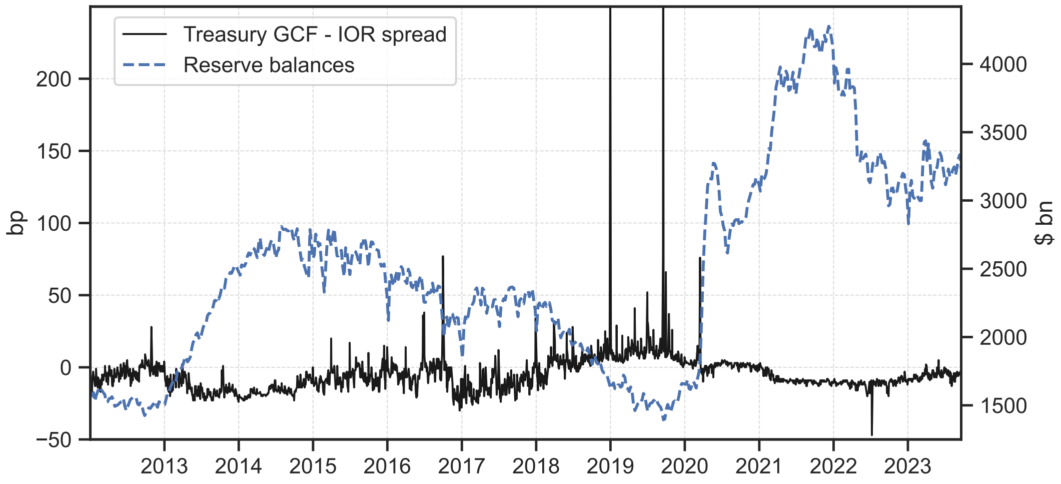

Figure 1: **Repo-IOR Spread and Reserve Balances**. The figure shows the series of weekly average reserve balances with federal reserve banks and the DTCC GCF Repo Index, which tracks the average daily interest rate paid for the most-traded GCF Repo contracts for U.S. Treasury minus the interest rate paid on reserves (IOR). The 2019 Sept. repo rate spike happened when reserve balances reached their 10-year lowest point.*Source: Federal Reserve Economic Data and Depository Trust & Clearing Corporation*

As the famous quote by Bernanke goes, ”The problem with QE is it works in practice, but it doesn't work in theory,” and so does QT. In a standard model, QT switches bank reserves to bonds one-for-one without any price impact, and thus, its effect does not ripple to the real side. It also seems hard to believe that the banking system, rid of financial stress, still needs so many reserves. The heart of my model is new precautionary regulations on liquidity coverage and leverage ratio proposed in Basel III, abiding by which banks adjust their assets and liabilities in response to QE/QT. In this sense, this model precisely intends to study the effect of QE/QT in the current ample reserve regime under new regulations with an emphasis on the exit strategy rather than to study the effectiveness of the initial QE after the GFC. The following three regulations established after the GFC are crucial to the model.

First, banks' ability to take on leverage is subject to capital requirements such as the *supplementary leverage ratio* (SLR). SLR is the US implementation of the Basel III Tier 1 leverage ratio. US banks are required to hold at least 3% common equity capital relative to their leverage exposure. The larger and systemic banks are subject to more strict requirements. As a result, large dealer banks face higher balance sheet costs and charge higher intermediation spreads. Moreover, banks are required to hold a 2% buffer above the 3% requirement. For these two reasons, I model the regulation as a leverage cost instead of a hard constraint of the leverage ratio, such that the bank leverage ratio fluctuates around the required level, and a tighter ratio implies a higher intermediation spread defined as the gap between bank asset return and deposit rate.

Second, according to Basel III, banks must hold high-quality liquid assets in potential preparation for cash outflows, referred to as *Liquidity Coverage Ratio* (LCR). Compared to other liquidity ratio requirements, LCR is more conservative in that it does not require a liquidity buffer for the very short term but looks further into future systematic liquidity stress. Under LCR, the minimum stock of unencumbered high-quality liquid assets (HQLAs) that banks must hold equals 100% of the net outflow over a 30-day stress period. Both reserves and treasuries are counted as tier 1 assets with no haircut. Other assets, such as government-backed MBS, are counted towards HQLAs with haircuts up to 40%. The regulation was proposed in 2010, but the 100% minimum was not enforced until 2019. In my model, I assume that the outflow needed to be backed is a fraction of the deposits that banks issue. Thus, the banks are required to hold reserves and bonds, adding up to a fraction of the total deposits.

Third, the repo market failure that triggered the GFC inspires the *Liquidity Stress Testing* (LST) requirement. Unlike LCR regulation that essentially requires HQLAs to back liabilities, LST establishes a constraint on the relative positions within HQLAs. As a supplement to LCR, LST instead focuses on large intra-day outflows (even though the daily net outflow is zero), which could still be sizable for large clearing banks or broker-dealers and cause stress. These outflows do not correspond to deposit outflows but are more likely triggered by the difference in timing of large repo and reverse repo transactions. LST thus requires enough reserves to back these outflows. Despite repo assets being nearly as liquid as reserves, banks demand enough reserves to back repo holdings under LST. In reality, a large fraction of these repos serve to finance bond holdings by shadow banks such as MMF and hedge funds. Therefore, LST effectively set up a cost for the financial system to hold treasury bonds, even though they are regarded as tier I assets in HQLAs. In my model, I collapse the bank and non-bank sectors into one financial sector. The way I model LST is to impose a cost of holding treasury that's a convex increasing function of the treasury to reserve ratio.

In the model, the central bank purchases the only long-lived asset, the perpetual treasury coupon bond, by issuing reserves. The reverse process, namely QT, means that the central bank sells the bonds for reserves. The LST cost-effectively set up a very inelastic demand for bonds for the banks, leading to a segmented bond market. As a result, the bank does not hold all the bonds that the Fed sells in QT but sells them to households that require a higher premium to accommodate a larger bond position, which depresses bond prices. This movement of relative holding of outstanding treasury bonds is consistent with the data as shown in Figure 2. When the central bank reduces its balance sheet size during QT periods (2017-2019 and 2023), banks reduce bond holdings while non-banks absorb the supply. If the LCR constraint binds, the banks with fewer HQLAs are forced to reduce deposits. Further, banks suffer a capital loss due to bond price collapse. These two factors both contribute to a balance sheet shrinkage, leading to a sale of working capital and consequent output fall. QT has adverse effects through the bond price and liquidity reduction channel.

I find a counter-intuitive result if the LCR constraint is not binding in the first place. Under the same QT path, when the Fed reduces liquidity supply to the bank, it does not have to shrink its balance sheet as the LCR constraint is loose. Instead, the banks are encouraged to hold more capital. QT effectively depresses the cost-adjusted financial asset return and has an expansionary real effect. Such effect ceases when the system hits the LCR constraint, indicating that the Optimal exit policy should be a rapid unwinding up to the LCR constraint followed by a very gradual further trajectory.

I use the model to study the optimal QE and its exiting strategy when adverse shocks of productivity, bond liquidity, and bank net worth hit the economy. As long as the economy dwells on the LCR constraint or is nearby in the steady state, a quick response of QE that provides liquidity and relaxes the constraint helps stabilize the real economy. An optimal exiting path features a slow QT whenever liquidity is scarce (LCR constraint is binding) and fast unwinding with overshooting when liquidity demand is satisfied.

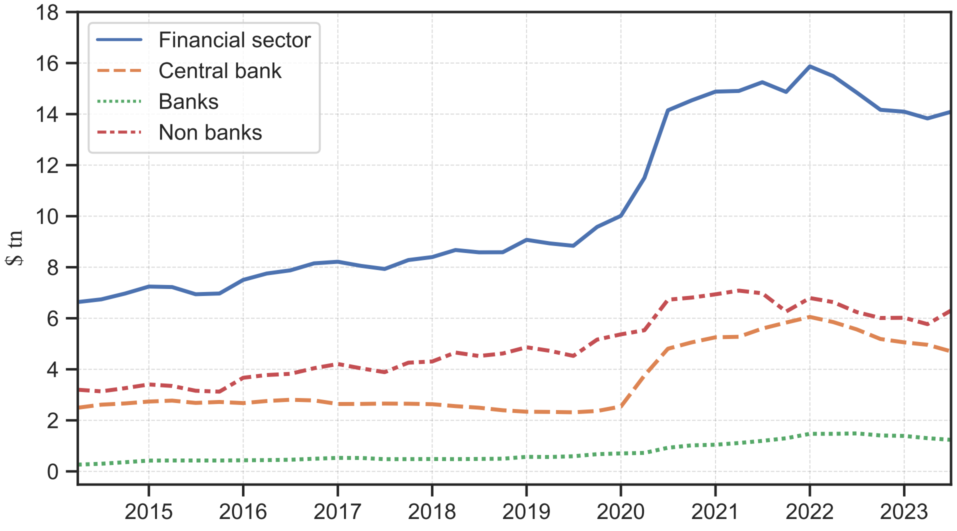

Figure 2: **Evolution of Treasury Holdings of Different Domestic Financial Institutions**. The figure shows that the central bank, banks, and non-bank financial institutions increase treasury holdings during QE periods. However, during the quantitative tightening period, the enlarged gap between treasury supply and the Fed's balance sheet is filled mainly with non-bank demand. Non-banks include pension funds, mutual funds, MMF, and insurance companies (I do not include hedge funds). The sample period is between Jan 2014 and Jun 2023.

2 Literature

This paper contributes to a growing literature that studies the channel through which quantitative easing works (restoring external financing Gertler and Karadi [2011], commitment device for low interest rate Bhattarai et al. [2023], lowering transaction cost Cui and Sterk [2021], etc.) and the effects of exiting in particular(Lopez-Salido and Vissing-Jorgensen [2023], Wei [2022]). Among many mechanisms proposed, I emphasize the role of segmented market (Chen et al. [2012], Vayanos and Vila [2021], Gourinchas et al. [2022]) and new liquidity regulations (d'Avernas and Vandeweyer [2020]) in driving scarcity of reserves and the movements of bond holdings and price. By analyzing the optimal QE/QT policies defined as the solution of a Ramsey problem, I add to the literature studying optimal asset-purchase programs (Curdia and Woodford [2011],Harrison [2017], Karadi and Nakov [2021]) and speak to some concerns arising from the use of QE/QT and the long-lasting sizeable central bank balance sheet in between studied in a list of papers (Woodford [2016], Acharya and Rajan [2022], Greenwood et al. [2016], Copeland et al. [2021], Diamond et al. [2020]).

My paper distinguishes itself from other papers of QE in the following ways. First, it allows occasional binding constraint instead of assuming the constraint permanently binds (such as incentive constraint in Gertler et al. [2012] and ZLB in Woodford [2016]). In these models, QE often only works on the constraint but has no effect when the constraint is not binding. On the contrary, QE/QT is always effective in my model, but mainly through a liquidity channel and in qualitatively very different ways depending on how close the economy is to the constraint. My paper is closely related to Karadi and Nakov [2021], which allows occasionally binding incentive constraint and concludes that an Optimal QT is gradual, but for different reasons. In their paper, QE responds to a negative net worth shock to the banks and thus needs to exit slowly to avoid damaging the bank's net worth and tightening the leverage constraint again. In my model, the effective margin is the liquidity coverage ratio constraint. Both of our papers indicate that QT off the constraint is expansionary. QT encourages capital accumulation in Karadi and Nakov [2021] because it raises asset returns. My model features an additional direct substitution from financial asset to working capital, thus implying an even more significant expansionary effect of QT off LCR constraint. Moreover, I model banks as bankers as utility optimizers with more refined balance sheet structures.

Second, I focus on the recently announced QT plan under new regulation frictions, which need to be addressed more in the existing literature. Third, I discuss the effectiveness of QE/QT with full commitment and from two very different perspectives. In a stationary economy, I study how a temporary asset-purchase program can help mitigate the impact of various shocks. In another case where QT is modeled as a unit-root process, I study the transitional path of the economy from a high balance sheet to a steady state with a smaller one and analyze the optimal unwinding policy.

I structure the rest of the paper as follows. I present the model in Section 3. I calibrate, solve the model, and analyze the results in Section 4. Section 5 discuss extensions and flaws. I conclude in Section 6.

3 Model

3.1 Environment

The model characterizes a dynamic two-sector economy populated by households and Bankers. Time is discrete and infinite. Bankers run banks facing regulation costs. The central bank controls the supply of reserves and bonds to the public. Quantitative tightening is modeled as the central bank selling treasury bonds in the market for reserves. Figure 3 illustrates the balance sheets of the three sectors.

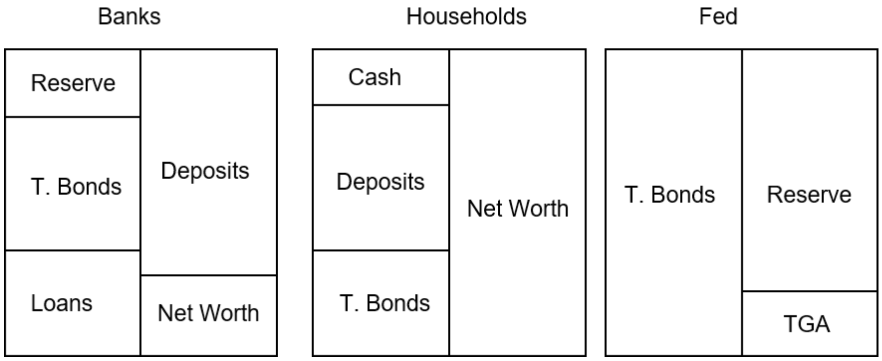

Figure 3: Agents' Balance Sheet

Market Structure Bankers hold reserves, bonds, and capital on the asset side. The central bank exogenously provides reserves with an interest rate paid, which is priced in equilibrium. Reserves are only used for inter-bank settlement, and the banking sector holds total reserves of $\bar{R}$. The bond in the model is a perpetual coupon bond paying unit each period. The market value of the bonds held by banks is qB^b. Banks hold K^b in the capital market with return r^k. Capital depreciates at rate δ. Banks take on leverage by issuing deposits D at a rate r^d. Household holds capital, deposit, and bond. Households' capital is less productive and an imperfect substitute for bank capital. Households have a desirable level of bond holding $\bar{B}_h$. Holding bonds in excess of this amount will incur an additional cost in the quadratic form with parameter χ. In equilibrium, the central bank would determine the total supply of bonds, and the capital demand is determined by the representative firm.

Preference and Technology Bankers and households have CRRA utility. Households value liquidity services provided by deposits in banks in a log form

$$U^h(C_t^h, D_t) = \frac{{C_t^h}^{1-\sigma}}{1-\sigma} + \gamma \log(D_t)$$

where γ governs the importance of liquidity services. In the extension, I allow a more general functional form of liquidity service affected by the households' position on other assets in the portfolio and the aggregate outcomes. Households' capital is less productive than banks' and is an imperfect substitute for the latter. The aggregate capital level is a CES combination of capital from the two sectors.

$$K = \left(K^{b\,\frac{\epsilon-1}{\epsilon}} + \theta K^{h\,\frac{\epsilon-1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1}}$$

For stationarity, bankers are less patient than households. The final good production adopts a Cobb-Douglas form with a unit labor endowment each period.

$$Y_t = A\exp(z_t)K_t^{\alpha}$$

The productivity z_t evolves according to a $AR(1)$ process

$$z_t = \rho_z z_{t-1} + \sigma_z \epsilon_t^z$$

Regulation Costs The bank faces three regulation costs. The first one is a standard *Leverage Ratio* (LR) constraint. Instead of a hard constraint, I assume that banks face a marginal deposit cost that is a function of their leverage ratio D/N^b. In the baseline model, any unit of deposit exceeding a fixed threshold LR induces a exponential loss with cost parameter κ and tightness parameter ζ in the following form.

$$g(D/N^b) = \text{cost}_{LR} = \frac{\kappa}{\zeta}\exp(\zeta(D/N^b - LR))$$

The second regulation cost is due to *Liquidity Stress Test* (LST) requirement. Banks are required to have enough reserves in preparation for intra-day deposit outflows before expected inflows arrive. Such intra-day reserve fluctuations typically happen during repo transactions with multiple parties at different times of the day. In reality, banks offer repos to shadow banks such as mutual funds and money market funds, which are major bond market participants. Now, if the banks are less able to provide repo services to the shadow banks, they have to raise deposits or repos from households that might require higher returns. In the baseline model, I collapse shadow banks with banks such that when the Fed tightens reserves, there is a decrease in bank-holding bonds and an increase in household-holding bonds. Households may require a higher bond yield due to the cost of holding bonds in excess of a desirable level. The bank's cost of holding bonds is determined by the ratio of bonds' market value to reserves. The functional form is exponential with a fixed threshold LST, cost parameter κ and tightness parameter ζ as follows

$$f(qB^b/R) = \text{cost}_{LST} = \frac{\eta}{\psi}\exp(\psi(qB^b/R - LST))$$

Finally, banks face an LCR constraint. The outflow in the stress period is proportional to the deposit, so the bank is required to back its deposits with HQLAs, namely reserves and bonds. Every unit of HQLA can back LCR units of deposits. In the steady state, bonds are as liquid as reserves. However, a liquidity shock could happen to the bond (repo) market. In such time, bonds will become less liquid. Though technically treasuries always count towards tier 0 HQLA, due to their dual role as repos in this model, banks that hold more repos face higher outflow/capital loss risk. An equivalent way to model the liquidity risk is to assume that the outflow risk (LCR requirement) when the shock hit is an increasing function of how much repo/bonds the banks hold. The constraint is

$$LCR(R + \omega qB^b) \geq D \tag{1}$$

where ω is the liquidity of bond subject to an AR(1) liquidity shock.

3.2 Agent's Problem

Bankers The problem written in sequential form is

$$\max_{\{C_t^b\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta_b^t \frac{{C_t^b}^{1-\sigma}}{1-\sigma}$$

subject to LCR constraint (1) and budget constraint

$$N_{t+1}^b = R_t^b(1 + r_{t+1}) + (q_{t+1} + 1)B_t^b - f(q_t B_t^b/R_t) + K_t^b(1 + r_t^k - \delta) - D_t^b(1 + r_{t+1}^d) - g(D_t/N_t^b) - C_t^b - \epsilon^N$$
$$N_t^b = R_{t+1}^b + q_t B_t^b + K_t^b - D_t^b$$

The variables r^k and r^d are the returns on capital and deposit. Bond return r^b is defined as $(q_{t+1} + 1)/q_t$. The bankers consume at the beginning of each period and then adjust their portfolios. Thus, the leverage ratio is calculated using the accounting net worth net of dividends. Bank is also subject to i.i.d. net worth shock ϵ^N.

Households Households maximize their lifetime utility discounted by β_h. Their problem written in sequential form is

$$\max_{\{C_t^h\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta_h^t U^h(C_t^h, D_t)$$

subject to their budget constraint

$$\begin{aligned} N_{t+1} &= K_t^h(1 + r_{t+1}^{k,h}) + D_t(1 + r_{t+1}^d) + (q_{t+1} + 1)B_t - \frac{\chi \bar{B}_h}{2} \left(\frac{B_t^h}{\bar{B}_h} - 1 \right)^2 + W_t - T_t \\ N_t &= K_t^h + D_t + q_t B_t^h \end{aligned}$$

The households' capital return $r_{t+1}^{k,h}$ is typically lower than the bank's capital return r^k due to lower productivity and relative abundance in equilibrium. Households have a unit labor endowment and earn a labor income of $W_t = (1 - \alpha)Y_t$.

Central Bank and Government The total supply of reserves $\bar{R}$ and total supply of bonds $\bar{B}$ is controlled by a central bank. When unwinding its balance sheet, the central bank reduces reserves by selling bonds according to

$$\bar{R}_t - \bar{R}_{t-1} = -q_t(\bar{B}_t - \bar{B}_{t-1})$$

Bond price q is determined in equilibrium, so at the start of each period, the change in bond price will incur a capital loss or gain to all of the three sectors. For the central bank, a bond value loss would lead to a lower remittance to the government. I assume that the government fully absorbs the fluctuation by adjusting the lump sum tax levied from households.

$$T_t = r_t \bar{R}_{t-1} + \bar{B}_{t-1} + G_t - (q_t - q_{t-1})\bar{B}_{t-1}$$

Government spending is assumed to be a constant G_{ss} over time. Note that the model does not assume a consolidated balance sheet. Though the central bank is the effective bond issuer, it can only adjust bond supply by reserve supply. Tax passively accommodates bond supply chosen by the central bank.

The optimal policy path for reserves is defined as the solution to the following Ramsey Problem. The central bank chooses the reserve path such that it minimizes a weighted sum of output gap and balance sheet cost [2]

$$\mathcal{L} = \min_{\{\bar{R}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta_h^t \left((Y - Y_{ss})^2 + \nu(\bar{R} - \bar{R}_{ss})\right)$$

3.3 Equilibrium

Definition Given an initial allocation of net worth $\{N_0^b, N_0^h\}$ and central bank policy path $\{\bar{R}_t, \bar{B}_t\}_{t=0}^{\infty}$, a sequential equilibrium is a set of stochastic processes for (i) exogenous variables $\{z_t, \epsilon^N, \omega_t\}_{t=0}^{\infty}$, (ii) prices $\{r_t, p_t, r_t^k, r_t^{k,h}, r_t^d\}_{t=0}^{\infty}$, (iii) control variables for banks and households $\{C_t^b, D_t^b, B_t^b, K_t^h, C_t^h, D_t^h, B_t^h, K_t^h\}_{t=0}^{\infty}$, (iv) banks and households' net worth $\{N_t^b, N_t^h\}_{t=0}^{\infty}$, and (v) aggregate outcomes $\{K_t, Y_t, W_t\}_{t=0}^{\infty}$ such that agents maximize their own problems and market clears

$$R_t^b = \bar{R}_t, \quad D_t^h = D_t^b, \quad B_t^h + B_t^b = \bar{B}_t, \quad Y_t = I_t^b + I_t^h + C_t^b + C_t^h + G_t + \text{costs}_t$$

4 Solution and Analysis

In this section, I derive the equilibrium asset price equations and show the channels through which QE/QT works. I calibrate the model to match key empirical moments. I first study the optimal bond purchase program and its exit as a whole. I shock the economy with negative changes in productivity, bond liquidity, and bank net worth and study the trajectory of reserve supply that stabilizes output while incurring limited balance sheet costs to the central bank. Then, I take a different view of QT as an unanticipated unit root process. I calibrate the baseline reserve trajectory along the transition path to the announcement of the current QT and how QT policies of different designs improve or impair efficiency given the targeted post-policy central bank balance sheet. I conduct these experiments from both the steady state with binding LCR constraint and one where the constraint is loose.

[2] I have the central bank optimize over output instead of agents' utility as there are two groups of agents with different preferences. In principle, the steady state is itself a trade-off between higher output and higher reserve interest cost, which is not hard in principle to implement. Instead, I let the central bank minimize deviation from the steady state for convenience.

4.1 Solving

In equilibrium, the bank's risk-adjusted expected return of each asset must be aligned. The gap between bond yield and capital return is the inconvenience yield due to the LST cost. The gap between capital return and deposit return is the balance sheet cost due to the LR cost.

$$\begin{aligned}
\mathbb{E}_t\left[\Lambda^b_{t+1}\frac{q_{t+1}+1}{q_t}\right] &= 1+\eta\exp\left(\psi\left(\frac{qB^b}{R^b_t}-LST\right)\right)-\lambda\omega * LCR\\
\mathbb{E}_t\left[\Lambda^b_{t+1}(1+r_{t+1})\right] &= 1-\eta\frac{qB^b}{R^b_t}\exp\left(\psi\left(\frac{qB^b}{R^b_t}-LST\right)\right)-\lambda * LCR\\
\mathbb{E}_t\left[\Lambda^b_{t+1}\left(1+r^d_{t+1}\right)\right] &= 1+\kappa\exp\left(\zeta\left(\frac{D_t}{N_t}-LR\right)\right)\\
\mathbb{E}_t\left[\Lambda^b_{t+1}(1+r^k_{t+1})\right] &= 1
\end{aligned}$$

The households are indifferent between holding capital, deposits, and bonds. The gap between capital return and deposit return is the liquidity services of deposits. Households require a higher return on bonds due to the frictional holding cost. Households' first-order conditions are

$$\begin{aligned}
\mathbb{E}_t\left[\Lambda^h_{t+1}(1+r^{k,h}_{t+1})\right] &= 1\\
\mathbb{E}_t\left[\Lambda^h_{t+1}(1+r^d)\right] &= 1+\frac{\lambda}{D}\\
\mathbb{E}_t\left[\Lambda^h_{t+1}\frac{1+q_{t+1}}{q_t}\right] &= 1+\chi\left(\frac{q_tB^h_t}{\bar{B}^h}-1\right)
\end{aligned}$$

4.2 Calibration

The calibration of the agent's preference and macroeconomic parameters are standard. Table 1 lists the parameter values. Time is quarterly, and Bankers are more impatient than households, targeting the steady-state size of the net worth of the banking sector relative to the households. Labor endowment is exogenous set to be one. I set the average productivity to normalize output to one. I set the quarterly depreciation rate δ to 0.05 to match a capital-output ratio of 5. For the financial side, I set the liquidity service γ to 0.05, targeting a zero real deposit rate. Households' desirable bond position $\bar{B}_h$ is set to 0.15 such that in equilibrium, households hold about 75% of bonds. The total supply of reserves and bonds is calibrated to the empirical level relative to GDP. Banks' balance sheet cost κ and tightness parameter ζ jointly decide the equilibrium leverage ratio and the elasticity of deposit to net worth shock. LST cost η and its tightness ψ also jointly target the elasticity of bank holding to reserve supply change. I will choose the LCR constraint according to whether the constraint should bind in the steady state. I set it to 4 for the efficient steady state, where the equilibrium LCR is 3.5. I set the central bank's balance sheet cost ν to be 0.01 in the baseline model. Results do not vary much as I lower ν to zero, but a higher ν tends to make the central bank less active.

In the unit-root QT analysis, the baseline process is calibrated to the announced length and pace. According to the Federal Open Market Committee, *"For Treasury securities, the cap will initially be set at $30 billion per month and after three months will increase to $60 billion per month."*, which means a quarterly decrease of about $120 billion worth of treasury bonds, about 5% of the current bank treasury holdings, and 4% of the current bank reserves. Though the Fed has not announced a stopping strategy, analysis has predicted that the current round of QT would last from 3-5 years, depending on the economic status. In the baseline QT, we start from a Reserve-to-bank-net-worth ratio analogous to the data and assume the central bank will reduce reserves by 4% for 20 quarters, resulting in ex-post reserves being 20% of the ex-ante level.

4.3 Benchmark Wihout LST Cost

Banks are willing to buy any amount of bonds at market price without LST Cost using reserves. The banks' arbitrage on the bond margin keeps the bond return and capital return equal. The only thing that changes after QT on the bank's balance sheet is a decrease in reserves and an increase in bonds by the same value. Consequently, no change happens to the households and other outcomes in the equilibrium. In fact, non-banks are the major participants in QE/QT operations, with large movements of bond positions observed on their balance sheets. However, The result does not rely on the fact that the central bank only trades with banks. In the model without LST regulations, suppose that households should withdraw their deposits for transactions. Then, the central bank clears the incoming deposits by reducing reserves by the same amount. Consequently, bond price tends to fall due to higher marginal holding cost, and deposit rate tends to rise due to scarcer liquidity service. In equilibrium, households will exchange all the bonds they bought from the central bank for deposits, and the price would be the same as the case where the central bank directly sells bonds to the bank. The results are also at odds with the relative movements of bond

Table 1: Calibration

Param'r	Value	Description	Target/Source
		A. Households	
β_h	0.99	Households' discount factor	Annual rate of 4%
$\bar{B}_h$	0.15	Desirable bond position	HH hold 75% bond outstanding
γ	0.05	Liquidity service	Real demandable deposit rate of 0%
χ	0.2	Bond holding cost	Elasticity of bond price to supply
σ	1	IES	
		B. Bankers	
β_b	0.95	Bankers' discount factor	Bank to HH net worth ratio of 10%
LR	3	Leverage ratio cost threshold	Bank leverage ratio of 5
ζ	1	Tightness of LR constraint	Deposit elasticity to net worth shock
κ	0.1	Leverage ratio cost level	Profit margin
η	0.1	LST cost level	
ψ	10	Tightness of LSR constraint	Elasticity of bank bond holding to reserve supply
LST	2	LST cost threshold	Equilibrium bank bond to reserve ratio of 2
LCR	4	Liquidity Coverage Ratio	Deposit to HQLAs ratio of 4
		C. Firm	
A	0.1	Average productivity	Normalize output to be unit
ϵ	3	ES of bank and HH capital	
θ	0.9	Household capital efficiency	Fraction of capital hold by banks
α	0.33	Capital share	labor share of 67%
δ	0.05	Depreciation rate	Capital output ratio of 5
		D. Government	
G_{ss}	0.2	Government spending	Government spending to GDP ratio of 25%
ν	0.01	Balance sheet cost	
$\bar{B}_{ss}$	0.2	Steady State private holding bond	Government outstanding to GDP ratio of 2
$\bar{R}_{ss}$	0.2	Steady State reserve level	Reserve to GDP ratio of 20%
		E. Exogenous process	
ρ_z	0.9	Persistence of productivity shock	
ρ_l	0.9	Persistence of liquidity shock	

positions of bank and non-banks observed in reality (See Figure 2).

4.4 Baseline Model Analysis

There are two ways to consider quantitative monetary policy through the lens of the model, depending on whether we believe the size of the central bank balance sheet is stationary and whether such policy is, to some extent, predictable. On the one hand, the Central bank balance sheet could be viewed as a mid-term monetary instrument to business and financial cycles that tends to revert to the long-run average after the shock diminishes. In this way, we expect the central bank to unwind quantitative easing at a certain time after the financial crisis, and the question is how long it will take and how clear the fed makes its policy trajectory. On the other hand, if there is no guarantee that the central bank balance sheet will return to its pre-shock size, then a permanent QE/QT shock, which is less anticipated than in the stationary economy, might have a qualitatively different effect on the economy.

In this part, I first take the view of a stationary economy and analyze how the economy responds to shocks and how QE might help mitigate the economic impact. Then, I relax the stationarity assumption and analyze an economy with a unit-root balance sheet process. Along with the baseline QT experiment calibrated to the Fed's announcement, I compare it with alternative trajectories that shrink the balance sheet of the same size in just one year.

4.4.1 Liquidity Shock

Starting from the steady state where bonds are as liquid as reserves, I shock the economy with a decline of bond liquidity to 0.5. Without monetary intervention, total liquidity drops, and LCR constraint binds. The bank has to decrease leverage by selling capital. Lower profitability reduces bank wealth temporarily. Due to the tight LCR constraints, bonds become more valuable as the only marketable liquid asset. The simultaneous decline of bond liquidity and price seems contradictory. The proper view consistent with the model is that some bonds lose their liquid asset property, but the safe liquid bonds gain more value. The overall outcome of the bond market turns out to be a boom. Banks still have to hold a great amount of bonds to back deposit, which incurs a large cost due to the liquidity stress test and depresses the bond market, offsetting the liquidity premium.

The optimal balance sheet response is quantitatively easing at the shock period and a prolonged unwinding period. As long as the LCR constraint binds, it is costly for banks to hold enough liquidity assets without help from the central bank. Thus, the optimal quantitative monetary policy needs to provide enough liquidity to the market until the shock vanishes and the market recovers. With monetary intervention, output falls by 0.27% initially compared to the 1.56% decline without QE. A faster unwinding of the initial balance sheet expansion will cause a sizeable initial output drop and slower recovery. There is no harm to the economy if the Fed maintains a large balance sheet for a more extended period, but under the balance sheet cost I chose, the Fed will only keep its size larger than the target level for 17 quarters.

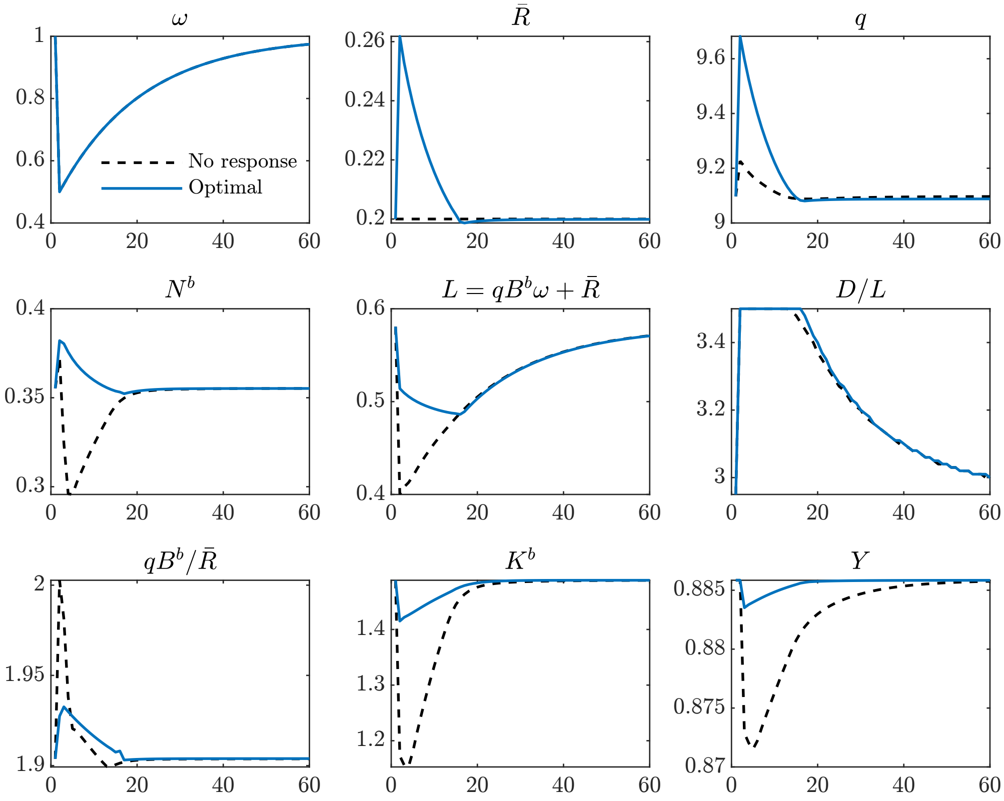

Figure 4: **IRF to an AR(1) Bond Liquidity Shock**. The figure shows the impulse response function of key economic variables to a AR(1) bond liquidity shock. Bond liquidity drops by 0.5 initially with quarterly persistence of 0.9. The economy starts from an non-binding steady state.

4.4.2 Productivity Shock

In some literature, quantitative easing is effective only in a narrow range of situations, for example, when the ZLB is binding. QE works in a broader range of situations in my model. In response to a 2% persistent productivity decline, banks suffer a loss. Lower net worth casts a higher cost on deposits and causes a capital sale. Total productive capital declines, exacerbating the output fall and output fall by more than 2%

The optimal monetary response is to infuse an additional 11.3% GDP worth of reserves into the banking system, followed by an expanded balance sheet for eight quarters. Note that the unwinding process overshoots and lasts much longer than QE. The reason is that QT is expansionary in this model, absent other shocks. Quantitative tightening forces the otherwise healthy banks to hold alternative assets, which can only be capital. Thus, when a negative shock happens, the optimal quantitative monetary policy features an initial QE to banks' intermediation capacity by restoring bank net worth or alleviating liquidity stress and a following QT process to incentivize banks to accumulate capital. This mechanism presents a different view of the speed of QT than Karadi and Nakov [2021] where QE helps restore banks' equity value but leads to a low return on the asset side, which slows down the bank's recovery. In my model, the central bank is not forced to slow down the unwinding process since the leverage constraint is not strictly binding. On the contrary, the central bank would like to unwind faster to encourage banks to sell liquid assets and invest more in capital. Thus, one caveat is that the unwinding process has to slow down if the recovering path hits the LCR constraint.

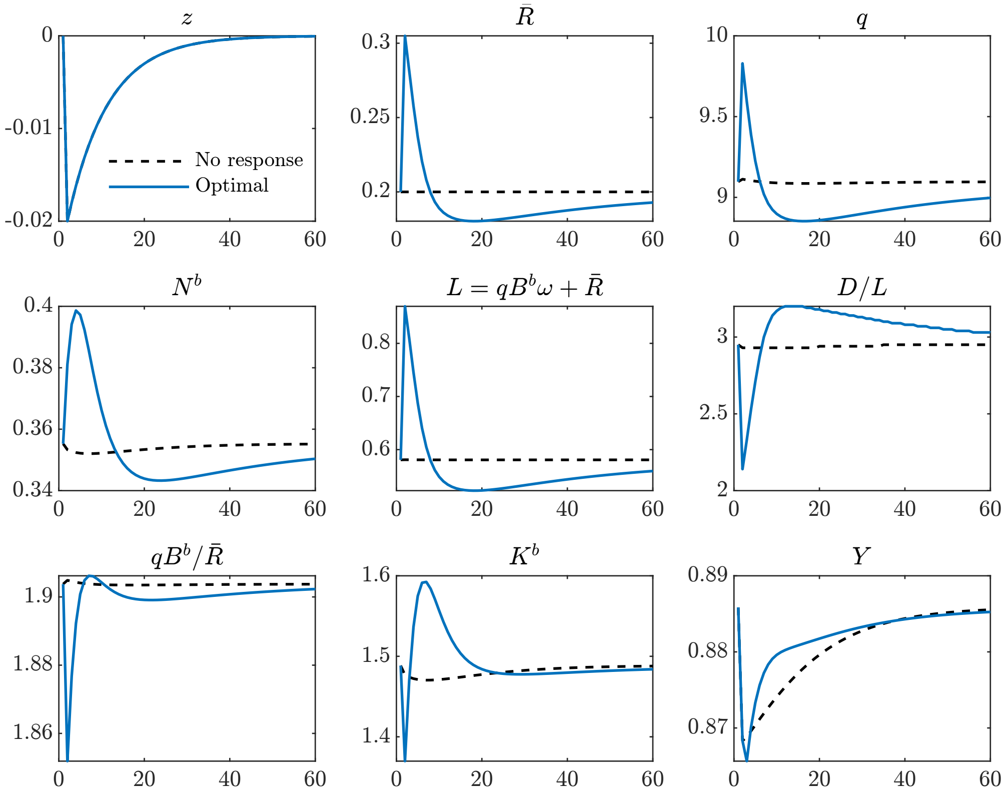

Figure 5: **IRF to an AR(1) Productivity Shock**. The figure shows the impulse response function of key economic variables to a AR(1) productivity shock. Productivity drop by 0.02% initially with quarterly persistence of 0.9. The economy starts from an non-binding steady state.

4.4.3 Financial Shock

I define a financial shock as an exogenous capital gain/loss by the bank. The economy's evolution hit by a financial shock is similar to a productivity shock. A financial shock destroys banks' net worth, forcing banks to decrease capital, which is amplified by a decrease in deposits due to a higher leverage ratio cost. QE inflates the bond market and restores the bank's net worth. Higher net worth helps banks to take leverage and accumulate capital faster. However, banks tend to substitute bonds for working capital when the central bank provides more liquidity and holding bonds is more lucrative. Thus, QE does not prevent a capital sell and output decline at the onset of the shock. Nevertheless, a larger balance sheet size later incentivizes the banks to accumulate capital faster in anticipation of future QT.

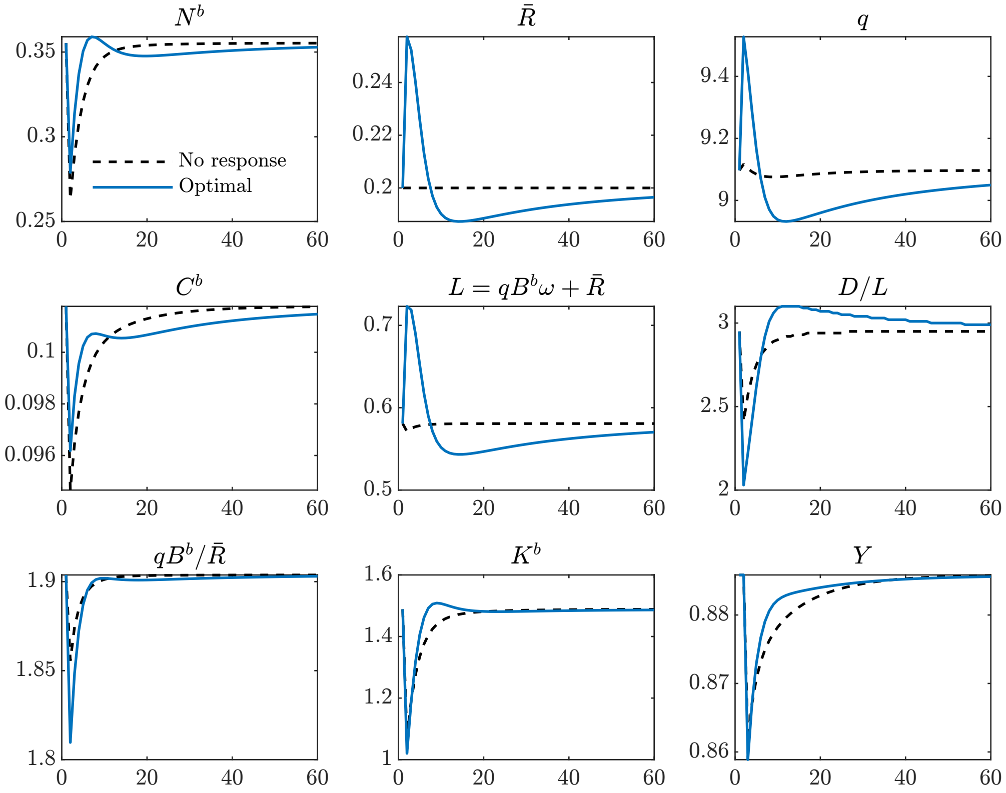

Figure 6: **IRF to an AR(1) Net Worth Shock**. The figure shows the impulse response function of key economic variables to an AR(1) bond liquidity shock with the central bank doing nothing or implementing QE/QT policy. Bank net worth drops by 33% initially. The economy starts from a non-binding steady state.

QE can prevent an output fall only if the steady state features a binding LCR constraint. The same decline in net worth now maps only one-to-one to disinvestment. The amplification from the leverage ratio margin is muted. Liquidity supply determines deposits issued. Consequently, QE helps banks take on leverage by providing more liquid assets. Households substitute deposits for capital, and banks invest more. The output fall is significantly milder. With a strictly binding LCR constraint, the optimal QT process becomes much more prolonged and never overshoots. Otherwise, a faster unwinding process will force banks to reduce leverage due to scarce liquidity too soon and hamper the recovery. The central bank has to wait for the private sector to accumulate capital, although supplying reserves incurs balance sheet costs and crowds out some investment.

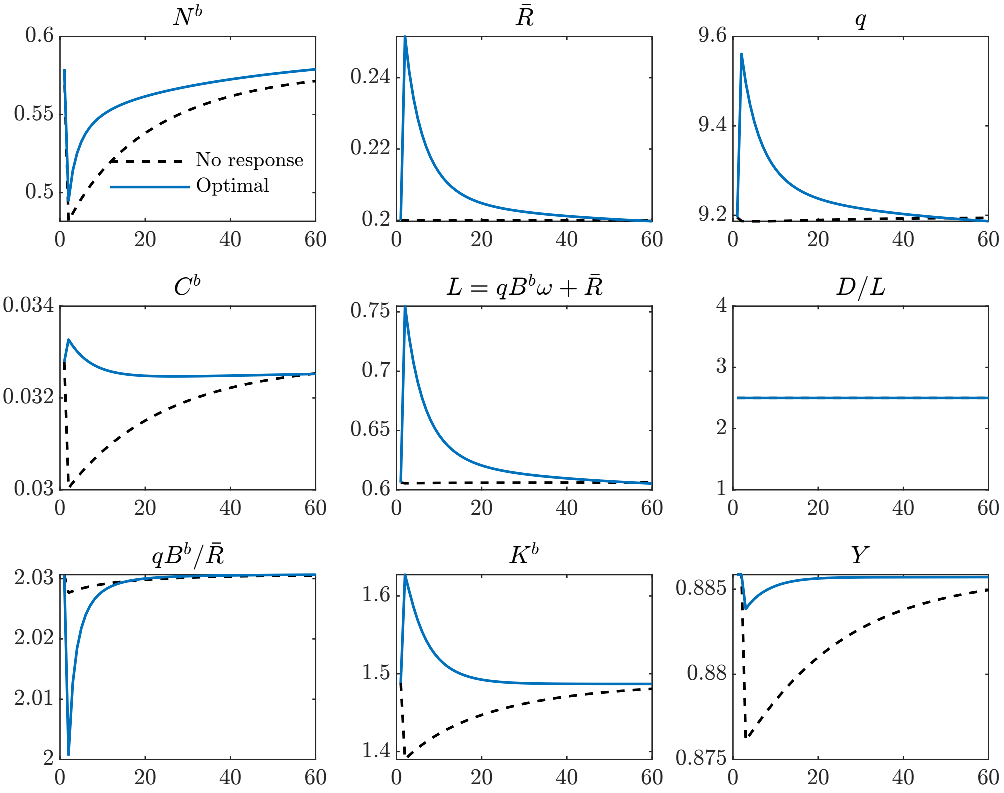

Figure 7: **IRF to an AR(1) Net Worth Shock (Binding LCR)**. The figure shows the impulse response function of key economic variables to an AR(1) bond liquidity shock with the central bank doing nothing or implementing QE/QT policy. Bank net worth drops by 33% initially. The economy starts from a steady state with a binding LCR constraint.

4.4.4 Non-stationary QE/QT

When the public believes the central bank balance sheet does not need to revert to its past level, the economy will converge to a new steady state with an expanded balance sheet. Suppose the Fed ”surprisingly” announces that it will start to unwind its balance sheet, and the public still believes that the policy indicates a permanent reserve supply change. In that case, the optimal QT must guide the economic transition to the new steady state.

Figure 8 shows the transition paths with a 33% permanent balance sheet reduction of different paces. As long as liquidity is not a concern, QT is expansionary. This counter-intuitive result is driven by the fact that the LCR constraint binds in the new steady state. Banks anticipate a limited leverage capacity and accumulate net worth by investing in capital. Although QT causes an initial decline in bank net worth by depressing the bond price, it does not map into less capital. Banks reduce consumption to accommodate the transition path to higher net worth instead. A slow

QT creates a longer boom by allowing higher deposit financing and faster capital accumulation. That said, a slow QT incurs balance sheet costs and pushes the boom further away, which gets discounted more. Note that if the steady state is close to the LCR constraint, then banks capital adjustment is limited and even turns negative when the buffer is too thin.

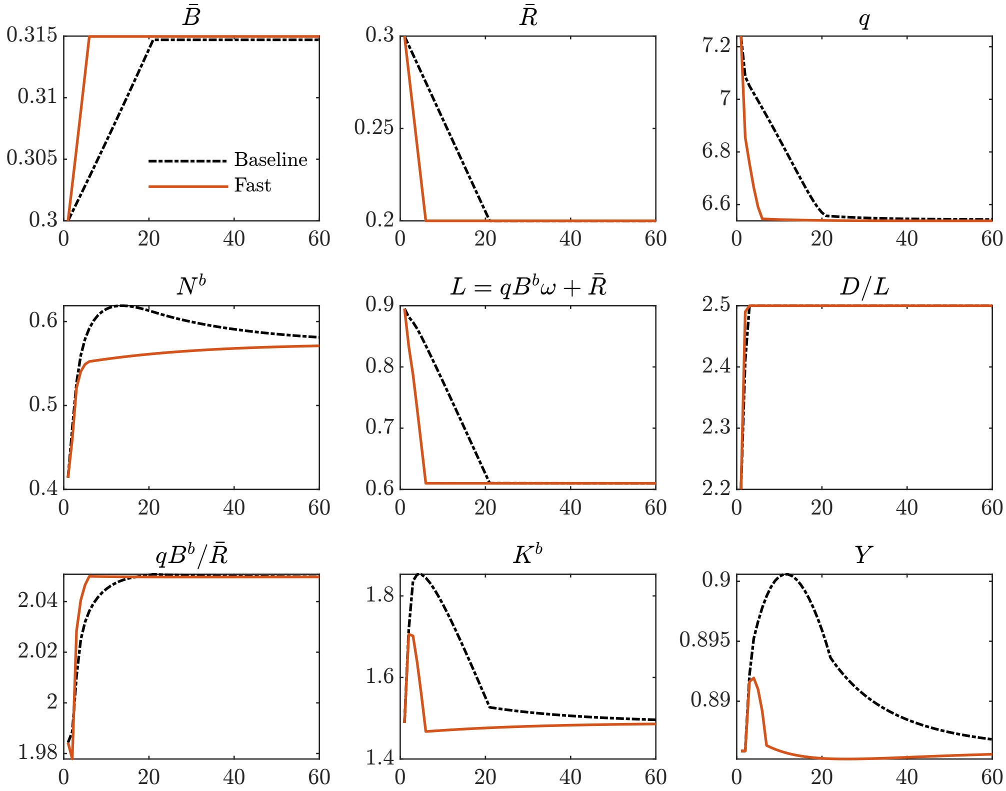

Figure 8: **IRF to an Unanticipated QT Path**. The figure shows the impulse response function of key economic variables to an unanticipated QT process. The economy start from an non-binding steady state.

The expansionary effect ceases to exist if the LCR constraint binds in the first place. Banks are forced to sell capital due to less supply of HQLAs. As Figure 9 shows, if the central bank reduces 33% of reserve unexpectedly within one year, the total HQLA supply falls by 33%, forcing banks to sell about 20% of their working capital. Output falls by 1.2% over eight quarters, followed by a slow recovery. Bond price will fall by 8%, leading to a 20% capital loss. Households require a higher premium on bonds because they need to hold bonds above the desirable level, which is partially mitigated by the more valuable liquidity service that bonds provide to banks. The initial wealth decline only translates one-for-one to capital decline because the leverage ratio margin is muted on the binding LCR constraint. Bankers cut consumption and accumulate net worth to sustain the desirable level of working capital (the same as the old steady state) in a new steady state with less external financing capacity. If the central bank gradually reduces the same amount of reserves in 5 years, then output falls gradually with a strictly higher trough. The severity of the recession depends on how tight the LST constraint is. The accumulation is slow due to the limited ability of deposit financing.

The optimal QT should be designed as follows. The central bank could tighten liquidity and drive the economy close to the LCR at a fast pace initially. When the LCR constraint binds, the trajectory becomes very gradual until the central bank achieves its goal. In reality, the exact number of the LCR constraint is not easily detected, so the central bank still needs to pace more slowly and detect the distance to the constraint or its tightness by closely watching the money market (spikes in the repo or bond market, for example). Studying the optimal exiting strategy under uncertainty of liquidity scarcity is a relevant next step for this research.

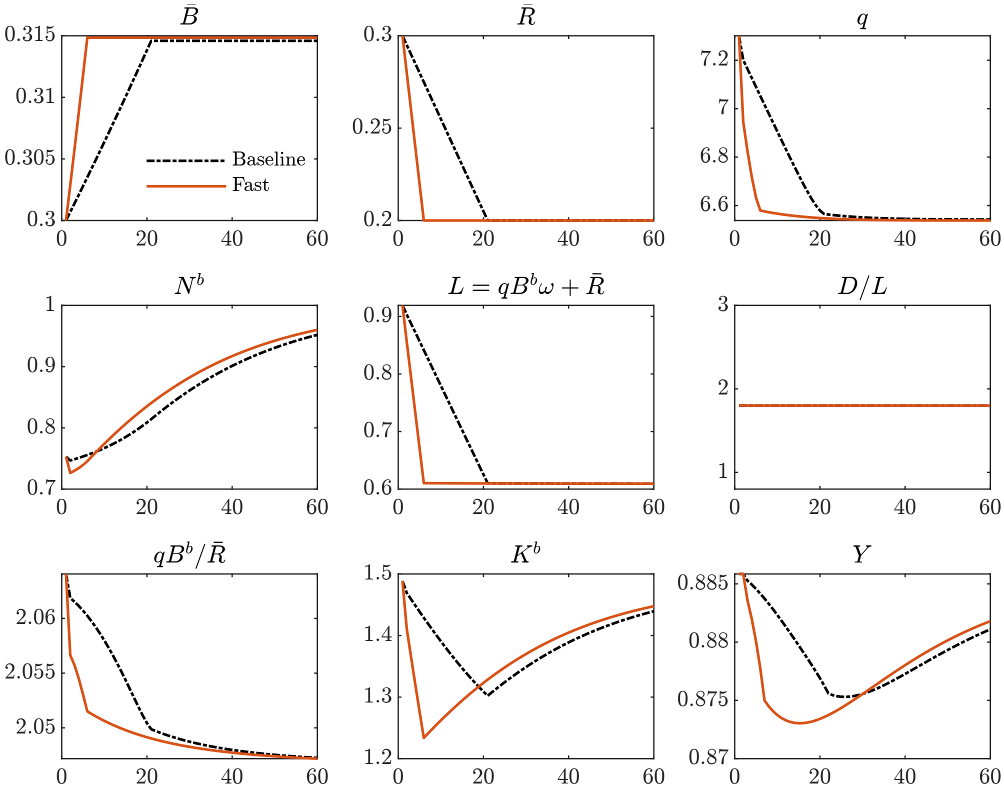

Figure 9: **IRF to an Unanticipated QT Path (Binding LCR)**. The figure shows the impulse response function of key economic variables to an unanticipated QT process. The economy starts from a steady state with binding LCR constraint.

5 Discussion and Extension

In this section, I include less formal discussions on the strengths and flaws of the model. I relate some of the model's properties to existing works and suggest possible extensions that will enable it to speak to a wider range of topics.

5.1 QE/QT and Interest Rate Policy

The model does not speak to the collaboration of interest rate policy and quantitative monetary policy, such as in Curdia and Woodford [2011]. It is nevertheless hard to talk about the effect of unconventional monetary policy on its own, as the Fed still regards interest rate policy as the first choice in response to conventional business cycles. Interest rate policies affect the cost of the balance sheet by changing the interest paid on reserves. It also directly changes asset prices, just as QT might potentially do. In this sense, the unit-root balance sheet analysis is a more proper way to understand the effect of QT perpendicular to other monetary policies. For the analysis of QE/QT as a monetary instrument for business cycles, it would make more sense to enrich the model with a New-Keynesian block and see whether QE/QT can add to the effectiveness of traditional monetary policies. But since QE/QT works mainly through the direct liquidity channel, it will remain effective with interest rate policies (whereas in Curdia and Woodford [2011], QE would not have any effect away from ZLB).

A related question is: why and when does the Fed choose interest rate hikes over QT? While both QT and interest rate hikes can raise long-term asset returns, their effects on investment might differ. While an interest rate hike makes holding bonds more attractive and crowds out investment, QT increases bond yield by generating a higher premium due to the segmented market and does not necessarily make investment less desirable. Through the lens of this model, the central bank will always first consider interest rate policy to mitigate shocks because QE/QT's effect depends on whether banks are close to the LCR constraint. Qt could even be expansionary when the banks have enough liquidity. Thus, an interest rate hike is a more consistent and effective policy instrument when the Fed wants to cool down the economy. The endeavor of unwinding the balance sheet should be regarded as a commitment to go back to normal and preparation of ammunition for the next crisis.

5.2 Deposit Finance and Liquidity Stress

As quantitative easing poured zero-maturity reserves into the banking system, the banks finance them with short-term liabilities. However, the co-movement of deposits and reserves is not symmetric with QT. Acharya et al. [2021] documented that the bank's liability duration does not increase [3] as the Fed shrinks reserves. Moreover, the banks keep issuing credit lines, which implies potential liquidity outflows just as other demandable liabilities. It seems that QT goes along with an increasing maturity mismatch that adds to liquidity risk.

My model explains the somewhat puzzling phenomenon. During QT, households hold more and more bonds and have a higher liquidity demand for liquidity services provided by deposits. As a result, the deposit rate falls relatively to other financial assets, and if LCR does not yet constrain banks, it will be willing to cater to the demand. On the contrary, during QE time, the LCR constraint might have been binding due to the initial financial shock. Thus, the deposit supply from the bank passively increased following the liquidity that QE restored. The asymmetry arises because QE is used in response to shocks while QT is conducted in a different economic environment.

Nevertheless, the caveat that QT might increase liquidity risk in Acharya and Rajan [2022] has its point. Deposit insensitive to reserve changes make it easier for the bank to hit the LCR constraint due to liquidity shock. In an ample reserve regime, it is hard to accurately determine where the threshold is and how much safety buffer is needed. These concerns argue for a gradual exiting plan but are absent in the deterministic model with MIT shocks. In the Appendix, I build a static model of QT with monopolistic-competitive banks that issue deposits and credit lines following Drechsler et al. [2017] and Li et al. [2019]. In the model, households are also allowed to hold money, and deposits are regarded as a less liquid and imperfect substitute for money. During QT, households demand more money but fewer deposits. Monopolistic banks increase deposit rates to attract deposits and further grant credit lines to business customers contingent on not withdrawing deposits. The sum of Deposit and credit lines increases when reserves decline. Incorporating this more enriched financial market structure into the dynamic model helps to study endogenous risk evolution along with UMPs.

[3] Demandable deposit actually decreases, but the magnitude is not comparable with the increase during QE of the same volume.

5.3 Steady State Efficiency

Instead of assuming a constraint that always binds, such as the incentive constraint in Gertler et al. [2012], I allow the model to start from an efficient steady state. Even if the steady state is close to the constraint, the dynamics can be qualitatively different. [4] In my model, a small buffer against the constraint greatly reverses the contractionary effect of QT. Gertler et al. [2012] assumes that the bank voluntarily chooses a balance sheet structure that is highly vulnerable to shocks that tighten the constraint. However, an occasionally binding constraint can provide more insights when uncertainty is involved, and we want to study how financial intermediaries' precautionary decisions change the robustness of the economy, such as in Brunnermeier and Sannikov [2014] where fire sale only occurs when shocks destroy experts' wealth beyond a certain level.

6 Conclusion

I analyze the optimal asset-purchase policy design, emphasizing exiting strategies using a model with banks facing balance sheet cost and liquidity ratio (LCR and LST) constraints. I study the optimal policy response to productivity, liquidity, and net worth shocks, starting from an efficient steady state and a constrained one. I also study the optimal unwinding policy modeled as a sequence of unit root shocks along a transition path to a new steady state with scarcer reserves. The central bank chooses the reserve path to minimize the weighted sum of the output gap and balance sheet cost. I found that QE is effective in response to all three of these shocks and is especially powerful when the LCR constraint binds by providing HQLAs to the banking system and restoring deposit financing. The optimal exiting path is gradual, especially as long as the LCR constraint binds, whether modeled together with QE or as an unexpected unit root shock. However, QT could be expansionary when liquidity is not a concern because higher bond-holding costs drive banks to substitute bonds for working capital. The interaction with interest rate policy and a more detailed financial market structure might be a fruitful avenue for future research.

[4] But not as different as presented in Figure 8 and Figure 9 because I keep the other parameters the same but adjust the constraint to have different steady states. This might be problematic in that the two economies have intrinsically different steady states and are not comparable. The correct approach is to re-calibrate parameters other than the constraint to generate steady states on the constraint and ones off but near the constraint.

References

V. V. Acharya and R. Rajan. Liquidity, liquidity everywhere, not a drop to use–why flooding banks with central bank reserves may not expand liquidity. Technical report, National Bureau of Economic Research, 2022.

V. V. Acharya, H. Almeida, F. Ippolito, and A. Perez-Orive. Credit lines and the liquidity insurance channel. *Journal of Money, Credit and Banking*, 53(5):901–938, 2021.

S. Bhattarai, G. B. Eggertsson, and B. Gafarov. Time consistency and duration of government debt: A model of quantitative easing. *The Review of Economic Studies*, 90(4):1759–1799, 2023.

M. K. Brunnermeier and Y. Sannikov. A macroeconomic model with a financial sector. *American Economic Review*, 104(2):379–421, 2014.

H. Chen, V. Cúrdia, and A. Ferrero. The macroeconomic effects of large-scale asset purchase programmes. *The economic journal*, 122(564):F289–F315, 2012.

A. Copeland, D. Duffie, and Y. Yang. Reserves were not so ample after all. Technical report, National Bureau of Economic Research, 2021.

W. Cui and V. Sterk. Quantitative easing with heterogeneous agents. *Journal of Monetary Economics*, 123:68–90, 2021.

V. Curdia and M. Woodford. The central-bank balance sheet as an instrument of monetarypolicy. *Journal of Monetary Economics*, 58(1):54–79, 2011.

W. Diamond, Z. Jiang, and Y. Ma. The reserve supply channel of unconventional monetary policy. *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*, 2020.

I. Drechsler, A. Savov, and P. Schnabl. The deposits channel of monetary policy. *The Quarterly Journal of Economics*, 132(4):1819–1876, 2017.

A. d'Avernas and Q. Vandeweyer. Intraday liquidity and money market dislocations. *Available at SSRN*, 2020.

M. Gertler and P. Karadi. A model of unconventional monetary policy. *Journal of monetary Economics*, 58(1):17–34, 2011.

M. Gertler, N. Kiyotaki, and A. Queralto. Financial crises, bank risk exposure and government financial policy. *Journal of monetary economics*, 59:S17–S34, 2012.

P.-O. Gourinchas, W. D. Ray, and D. Vayanos. A preferred-habitat model of term premia, exchange rates, and monetary policy spillovers. Technical report, National Bureau of Economic Research, 2022.

R. Greenwood, S. G. Hanson, J. C. Stein, et al. The federal reserve's balance sheet as a financial-stability tool, 2016.

R. Harrison. Optimal quantitative easing. 2017.

P. Karadi and A. Nakov. Effectiveness and addictiveness of quantitative easing. *Journal of Monetary Economics*, 117:1096–1117, 2021.

W. Li, Y. Ma, and Y. Zhao. The passthrough of treasury supply to bank deposit funding. *Columbia Business School Research Paper, USC Marshall School of Business Research Paper*, 2019.

D. Lopez-Salido and A. Vissing-Jorgensen. Reserve demand, interest rate control, and quantitative tightening. *Interest Rate Control, and Quantitative Tightening (February 27, 2023)*, 2023.

D. Vayanos and J.-L. Vila. A preferred-habitat model of the term structure of interest rates. *Econometrica*, 89(1):77–112, 2021.

B. Wei. Quantifying "quantitative tightening"(qt): How many rate hikes is qt equivalent to? 2022.

M. Woodford. Quantitative easing and financial stability. Technical report, National Bureau of Economic Research, 2016.

7 Appendix

7.1 A Model of QT and Bank Demandable Liability

To address the anomaly of the rise of deposits and credit lines during QT in Acharya and Rajan [2022], I build a static model where banks use credit line issuance to maintain customer relationships. Banks offer cheaper credit line services in QT time to dissuade deposit withdrawal. I borrow the framework of the liquidity demand model from Drechsler et al. [2017] and Li et al. [2019] and study how QE and QT affect band liability composition. The investor maximizes utility, which consist of the wealth level and liquidity value.

7.1.1 Model Setup

The market participants of the economy are representative investors and N banks that issue deposits that are imperfect substitutes and credit lines that are perfectly addable. To focus on the liability side, I assume that the bank can only issue loans at its chosen rate r^l on the asset side. Each bank i can determine a deposit spread $s_i = r - r_i^D$ that they charge their investors. The bank can commit to providing a credit line of c_i dollars for every dollar of deposit. The credit is drawn when the firm faces liquidity needs and is perceived as liquid as money. However, it only shows up on the balance sheet once it is used. The representative investor's total deposit is an aggregation of deposits in each banks, defined as

$$D = \left(\frac{1}{N} \sum_{i=1}^{N} D_i^{\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}}$$

Investor's Problem The investor can hold an illiquid asset that has a return of r. Due to a liquidity preference, she holds a portfolio of deposits, bonds, and money. Deposits and bonds are less liquid than money but have higher returns. The investor maximizes utility, a weighted sum of the wealth level and liquidity service.

$$W_1 + \theta \log(L)$$

Liquidity preference could be interpreted as transaction needs or preparation for liquidity needs in stress states. The liquidity value L is defined as

$$L = \left(\delta D^{\frac{\epsilon-1}{\epsilon}} + \beta B^{\frac{\epsilon-1}{\epsilon}} + M^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}}$$

where D, B, and M refer to deposit, bond, and money-like liquidity, respectively. The liquidity of money is normalized to one. Deposits and bonds have different degrees of liquidity δ and β and are imperfect substitutes for money. Thus, quantity matters for returns. Money has strictly zero return. Investors that put deposit D_i in bank i will be offered a credit line of size $c_i D_i$ where c_i is chosen by the bank. I do not intend to claim that the pre-committed credit is as liquid as money but rather make this assumption for simplicity. I will show that the result does not change qualitatively if the assumption is relaxed.

$$M = \bar{M} + \frac{1}{N}\sum_{i=1}^{N} c_i D_i$$

Denote the FFR-bond spread and deposit spread as l and s. Given initial value W_0, the budget constraint is

$$W_1 = W_0(1+r) - Ds - Bl - Mr$$

In the symmetric equilibrium, the deposit demand could be solved following a two-layer approach. I first solve the aggregate deposit demand and then solve the allocation of deposits which is formulated as the following

$$\max_{D_i} -\frac{1}{N}\sum D_i s_i + \theta \log(L)$$

The investor wants to hold deposits because it provides positive returns, but also because of the liquidity value of credit services linked with deposits.

Bank's Problem Bank face a loan demand function as a function of loan rates.

$$Q(r_i^l) = \bar{Q} - 2\gamma r_i^l$$

Banks have market power over firms with limited financial access and can set loan rates. The total loan supply must equal the deposit the bank issues. The bank's profits depend on the loan spread, deposit spread, how many credit lines the bank issues, and the issuing cost. The bank's optimization problem is

$$\max_{s_i,c_i}\left(s_i+\frac{\bar{Q}}{2\gamma}-r-\frac{D_i}{2\gamma}-\kappa c_i\right)D_i$$

where c is the credit line service the bank promises to the client in the sense that if the investor deposit in bank i with D_i, then the bank issue cD_i amount of credit line to the investor.

7.1.2 Model Solution

The FOCs are

$$s=\frac{\theta}{L}\delta\left(\frac{D}{L}\right)^{-\frac{1}{\epsilon}}\quad l=\frac{\theta}{L}\beta\left(\frac{B}{L}\right)^{-\frac{1}{\epsilon}}\quad r=\frac{\theta}{L}\left(\frac{M}{L}\right)^{-\frac{1}{\epsilon}}$$

From which we can get

$$L=D\left[\delta+\beta^{\epsilon}\delta^{1-\epsilon}\left(\frac{s}{l}\right)^{\epsilon-1}+\delta^{1-\epsilon}\left(\frac{s}{r}\right)^{\epsilon-1}\right]^{\frac{\epsilon}{\epsilon-1}}$$

and the expression for deposit demand

$$D=\frac{\theta}{s}\left[1+\left(\frac{\beta}{\delta}\right)\left(\frac{\beta s}{\delta l}\right)^{\epsilon-1}+\left(\frac{1}{\delta}\right)\left(\frac{s}{\delta r}\right)^{\epsilon-1}\right]^{-1}$$

Thus

$$B=\left(\frac{s\beta}{l\delta}\right)^{\epsilon}D=\frac{\theta}{l}\left[1+\left(\frac{\delta}{\beta}\right)\left(\frac{\delta l}{\beta s}\right)^{\epsilon-1}+\left(\frac{1}{\beta}\right)\left(\frac{l}{\beta r}\right)^{\epsilon-1}\right]^{-1}$$

The first order condition of the deposit allocation problem is

$$-\sigma_i=\mu\left(\frac{D_i}{D}\right)^{-\frac{1}{\eta}}$$

where

$$\sigma_i=s_i-\tilde{c}_i=s_i-\theta\frac{1}{L}\left(\frac{M}{L}\right)^{-\frac{1}{\epsilon}}c_i<s_i$$

where $\tilde{c}_i$ is liquidity-adjusted credit line service. Thus

$$-\frac{\partial\log D_i}{\partial\log s_i}=\frac{s_i}{\sigma}\left(1-\frac{1}{N}\right)\eta,\quad\frac{\partial\log D_i}{\partial\log c_i}=\theta\frac{1}{L}\left(\frac{M}{L}\right)^{-\frac{1}{\epsilon}}\frac{c_i}{\sigma}\left(1-\frac{1}{N}\right)\eta$$

We can see that the demand elasticity to credit line service is higher when money is in shortage, providing the bank a stronger incentive to provide liquidity via credit line. Let $\mathcal{M} = (1 - 1/N)\eta$, in equilibrium the following two equations must hold

$$\frac{D_i}{\gamma} = \frac{\bar{Q}}{2\gamma} - r - \kappa c_i + s_i \left(1 - \frac{\sigma_i}{s_i \mathcal{M}}\right)$$
$$\frac{D_i}{\gamma} = \frac{\bar{Q}}{2\gamma} - r + s_i - \kappa c_i \left(1 + \frac{\sigma_i}{\tilde{c}_i \mathcal{M}}\right)$$

The solution is

$$\kappa = \theta \frac{1}{L} \left(\frac{M}{L}\right)^{-\epsilon}$$

The equilibrium is characterized by

$$D(s,l,r,c) = \hat{D}(s,r,c)$$
$$B(s,l,r,c) = \bar{B}$$
$$M(s,l,r,c) = \bar{R} + cD$$

QT is an operation that reduces $\bar{R}$ by increasing $\bar{B}$, keeping the sum still. In equilibrium, total liquidity falls, and investors withdraw deposits for liquidity. When money is in a scarcer supply, banks put more effort into providing liquidity service by issuing more credit lines contingent on fewer deposit withdrawals. Total demandable liability $(1 + c)D$ increases after QT as it can be proved that for parameters in a domain the elasticity of credit line supply exceeds the elasticity of deposit demand.

$$\frac{\partial c}{\partial \bar{R}} \frac{\bar{R}}{1+c} > -\frac{\partial D}{\partial \bar{R}} \frac{\bar{R}}{D}$$